EXHIBIT AW

Santa Clara University School of Law

Santa Clara Law Review

Volume 43 | Number 1 Article 3

1-1-2002

How the SAT Creates Built-in-Headwinds: An Educational and Legal Analysis of Disparate Impact

William C. Kidder

Jay Rosner

HOW THE SAT CREATES "BUILT-IN HEADWINDS": AN EDUCATIONAL AND LEGAL ANALYSIS OF DISPARATE IMPACT

William C. Kidder* & Jay Rosner**

> With the end of affirmative action, it is more apparent than ever that the old-time preferences for folks who are privileged by race and class have never died.
>
> —Charles R. Lawrence III[1]

---

* Researcher, Testing for the Public, Berkeley, California. J.D., Boalt Hall School of Law, University of California, Berkeley; B.A., University of California, Berkeley. Mr. Kidder served as a consultant for the student intervenors defending affirmative action in *Grutter v. Bollinger*, 137 F. Supp. 2d 821 (E.D. Mich. 2001), *rev'd en banc* 288 F.3d 732 (6th Cir. 2002). He also conducted research on affirmative action for the Society of American Law Teachers (SALT), and subsequent to the acceptance of this article took a position as Law Clerk to the Honorable Edward M. Chen, Northern District of California. The authors wish to thank the following scholars for their helpful reviews: Vikram Amar, Angelo Ancheta, Richard Delgado, Lani Guinier, William Henderson, Bradford Mank, David Oppenheimer, Peter Sacks, Gerry Spann, Sam Spital, and David M. White. We also appreciate the assistance of Martin Shapiro in transforming our SAT data into a useable form. Mr. Kidder also thanks Gale Drake Jones for her support and patience with this project.

** Executive Director, The Princeton Review Foundation. J.D., Widener University; B.A., University of Pennsylvania. Member of the New York, New Jersey, Pennsylvania, and Oregon Bars. Mr. Rosner recently testified as an expert witness on standardized testing for pro-affirmative action student interveners in *Grutter v. Bollinger*. Mr. Rosner's expert report is reprinted in 12 LA RAZA L.J. 377 (2001).

1. Charles R. Lawrence III, *Two Views of the River: A Critique of the Liberal Defense of Affirmative Action*, 101 COLUM. L. REV. 928, 943 (2001).

I. INTRODUCTION

In *Griggs v. Duke Power Co.*,[1] the Supreme Court declared that Title VII of the 1964 Civil Rights Act extends to acts of unintentional discrimination.[2] The Court held that Duke Power's reliance on graduation requirements and standardized test scores as hiring and transfer criteria violated Title VII because these requirements invidiously discriminated against African Americans and yet were unrelated to actual job performance.[3] *Griggs* was the case of first impression in which the Court established a framework for assessing "disparate impact" discrimination, criticizing the unwarranted reliance on standardized tests that operate as "built-in headwinds" against minority groups.[4]

This article analyzes the SAT's disparate impact, and demonstrates how "built-in headwinds" are designed into the actual process of selecting and developing SAT questions.[5] Although this process may appear facially-neutral and non-discriminatory, the SAT unfairly exacerbates the test's already significant disparate impact on African Americans and Chicano test-takers.[6] Part I provides an

---

1. Griggs v. Duke Power Co., 401 U.S. 424 (1971).

2. *See id.* We note that the Court was generous in characterizing the employer's policy as "unintentional discrimination." Prior to 1965, the Duke Power Company's Dan River plant in North Carolina explicitly discriminated against African Americans; it was no coincidence that the new diploma/test score hiring criteria were first imposed on July 2, 1965, the very day that the Civil Rights Act took effect. *See id.* at 426-27.

3. *See id.* at 431-36.

4. *See id.* at 432.

5. *See* discussion *infra* Parts II and III.

6. A few points about race and ethnicity. First, throughout the text we capitalize "White" and "Black" intentionally to designate these as specific groups deserving of proper noun status because these categories have deep political and social meanings. While there is some disagreement among scholars, this capitalization is consistent with many critical race theorists. *See, e.g.*, IAN HANEY LOPEZ, WHITE BY LAW xiv (1996); Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1707, 1710 n.3 (1993).

Second, in the interest of accuracy, this article uses both the terms Chicano and Latino. The data on 1998 SAT questions in Part II refers to Chicanos (Mexican Americans) because the data we obtained from ETS reported Chicanos separately. However, most of Parts III and IV refer to Latinos because the authors of the studies we discuss report data on Latinos (which includes Chicanos, as well as those with national origins in Central America, Cuba, Puerto Rico, and South America). Since Latino is a broad ethnic category referring to all people of Hispanic origin (which can include people who self-identify their race as either White or Black), when we refer to White we mean non-Hispanic White, and when we refer to Black we mean non-Hispanic Black. All of these categories are the subject of enduring debate, and more recently the literature is divided on whether to use Chicano or Chicana/o and Latino or Latina/o. *See, e.g.*, Margaret E. Montoya, *A Brief History of Chicana/o School Segregation: One Rationale for Affirmative Action*, 12 LA RAZA L.J. 159 (2001); Ian F. Haney Lopez, *Protest, Repression, and Race: Legal Violence and the Chicano Movement*, 150 U. PA. L. REV. 205, 208 (2001); Rachel F. Moran, *What if Latinos Really Mattered in the Public Policy Debate?*, 85 CAL. L. REV. 1315 (1997); Francisco Valdes, *Poised at the*

overview of standardized tests in relation to recent affirmative action litigation and admissions policy changes. This article challenges the conventional wisdom that the SAT accurately measures merit and fairly reflects group differences in educational attainment.

Parts II and III provide evidence of the existence of racial and ethnic bias on the current SAT. Part II analyzes previously undisclosed data about SAT questions and demonstrates that a substantially higher proportion of White test-takers correctly answer virtually all questions on the scored sections of the SAT. Contributing to the larger educational debate about test bias, reliability, and construct validity in Part III, our findings indicate that the test development process unintentionally, but consistently and predictably increases the disparate impact of the SAT. Moreover, traditional methods of rooting out biased questions are ineffective and are based on a dubious premise. A more effective method to lessen disparate impact can be achieved by means of *Golden Rule*-style techniques for selecting test questions. This part also addresses criticisms of our proposed bias reduction method and some practical difficulties in implementation of this method.

Part IV provides a detailed analysis of the law governing standardized tests, university admissions, and Title VI disparate impact claims. The prospect of enforcing U.S. Department of Education disparate impact regulations through section 1983 is still a viable option, notwithstanding many difficulties. Another possible but not as

---

Cusp: LatCrit Theory, Outsider Jurisprudence and Latina/o Self-Empowerment, 2 HARV. LATINO L. REV. 1, 2 n.1 (1997).

Third, we limited our study to African Americans and Chicanos for a combination of policy and empirical reasons. These groups have been hardest hit by the end of affirmative action in higher education. We were not able to analyze the disparate impact of SAT items on American Indians because of the small absolute number of American Indians who take the SAT. For background on American Indians, affirmative action, and educational access, see Carole Goldberg, *American Indians and "Preferential" Treatment*, 49 UCLA L. REV. 943 (2002); Faith Smith, Expert Report in Grutter v. Bollinger, 137 F. Supp. 2d 821 (E.D. Mich. 2001) (no. 97-75928), *reprinted as Building Native American Representation in the Law: The Need for Affirmative Action, in* 12 LA RAZA L.J. 397 (2001); Gloria Valencia-Weber, *Law School Training of American Indians as Legal-Warriors*, 20 AM. INDIAN L. REV. 5 (1995-96). With our database from ETS it was also not possible to separately analyze Asian Pacific American subgroups, nor were we able to combine Chicanos with other Latinos. Thus, we were unable to look at Asian Pacific American groups that tend to be under-represented in higher education, including Filipinos and Southeast Asians. However, in Part III.D. we comment on how the application of impact reduction techniques for African Americans and Latinos might effect women and Asian Pacific Americans overall.

favorable alternative is to file a complaint with the Office for Civil Rights. While the evidence in Parts II and III focuses on racial/ethnic bias in the SAT test construction process, independent of this evidence it remains the case that many universities are vulnerable to disparate impact challenges over their use of the SAT for reasons discussed in Part IV. Many universities may not be able to meet their "educational necessity" burden because they knowingly use the SAT in ways that have not been validated, as is the case with rigid cut-off scores. Even more institutions may have difficulty establishing educational necessity because the SAT only incrementally improves the prediction of college grades, and is even less helpful in forecasting graduation rates or contributing to colleges' institutional goals. Even when educational necessity is established, a plaintiff in cases challenging use of the SAT may still prevail by demonstrating that percentage plans or SAT-optional admissions are less discriminatory alternatives that are equally effective in meeting the educational goals of a university.

Part V discusses the consequences of ending affirmative action at public universities in California, Georgia, and Texas. While it remains unclear whether the Supreme Court will ultimately uphold higher education affirmative action programs, either way there are steps that can be taken in the test construction process to lessen the SAT's disparate impact on African Americans and Chicanos without compromising the test's validity. Since test producers have been extremely resistant to the kinds of test development changes advocated in this article, we conclude that ending reliance on the SAT, or making the test optional, may be the most pragmatic strategies for fostering equity and fairness in university admissions. This article's purpose is to document the ways in which the current SAT development process amounts to covert racial gerrymandering in favor of Whites, thereby exacerbating disparate impact against students of color.

A. *The SAT and Affirmative Action*

The SAT has long been the gatekeeper of higher education.[7] Among the 2.85 million American high school graduates in 2001, 1.3 million took the SAT, and over half of those took the test two or more

7. Former College Board President George Hanford states that "the SAT served as the most widely used and possibly the most important single talent search device the country had." GEORGE H. HANFORD, LIFE WITH THE SAT: ASSESSING OUR YOUNG PEOPLE AND OUR TIMES 90 (1991).

times.[8] In addition, 1.1 million high school students, predominantly in the Midwest and the South, took the ACT, the only alternative college admissions test to the SAT.[9] In the last two decades, the proportion of high school graduates taking the SAT grew from 33% to 45%.[10] The College Board, as owner of the SAT, and the Educational Testing Service (ETS), as administrator and designer of the test, last year combined to take in $900 million in gross revenue.[11]

In 2001, over one-third of all SAT test-takers were students of color, an all-time record.[12] Yet at the same time, opponents of affirmative action mounted a spirited, multi-faceted, and often successful attack on race-conscious college admissions. As a consequence, public universities discontinued race-conscious admissions in Texas,[13] California,[14] Florida,[15] Washington,[16] Georgia,[17]

---

8. *See* Press Release, College Board, 2001 College Bound Seniors Are the Largest, Most Diverse Group in History (2001) [hereinafter College Board Press release]; Ben Gose & Jeffrey Selingo, *The SAT's Greatest Test*, CHRON. HIGHER EDUC., Oct. 26, 2001, at A10.

9. *See* Ben Gose, *ACT Sees Openings for Expansion in Debate Over the SAT*, CHRON. HIGHER EDUC., Oct. 26, 2001, at A13. Note that a small proportion of students take both the SAT and the ACT.

10. *See* Gose & Selingo, *supra* note 8, at A10.

11. *See id.*

12. *See* College Board Press Release, *supra* note 9.

13. *See* Hopwood v. Texas, 78 F.3d 932 (5th Cir. 1996) (ruling that the affirmative action program at the University of Texas (UT) Law School was unconstitutional because taking account of race to improve the quality of learning was not a compelling governmental interest, and because the program was not narrowly tailored to remedy the UT Law School's own prior discrimination against minority students); Hopwood v. Texas, 236 F.3d 256 (5th Cir. 2000); Chris Vaughn, *Order Lifted in College Entry Case, Court Maintains Ban on Race-based Admissions*, FORT WORTH STAR-TELEGRAM, Dec. 22, 2000, at 1.

14. Proposition 209, now CAL. CONST. art. I, § 31, was passed by a 54% majority of California's voters in November 1996. It states: "The State shall not discriminate against, or grant preferential treatment to, any individual or group on the basis of race, sex, color, ethnicity, or national origin in the operation of public employment, public education, or public contracting." CAL. CONST. art. I, § 31. For a description of the political fight over Proposition 209, see LYDIA CHAVEZ, THE COLOR BIND: CALIFORNIA'S BATTLE TO END AFFIRMATIVE ACTION (1998). Civil rights organizations mounted an unsuccessful challenge to the constitutionality of Prop. 209. *See also* Coalition for Economic Equity v. Wilson, 110 F.3d 1431 (9th Cir. 1997).

In addition, the University of California (UC) Regents approved the SP-1 Resolution in July 1995. SP-1 ended race-conscious admissions at the graduate and professional level beginning on January 1, 1997, a year before Proposition 209 took effect. *See* Kit Lively, *University of California Ends Race-Based Hirings, Admissions*, CHRON. HIGHER EDUC., July 28, 1995, at A27; William C. Kidder, *Situating Asian Pacific Americans in the Law School Affirmative Action Debate: Empirical Facts About Thernstrom's Rhetorical Acts*, 7 ASIAN L.J. 29, 34-35 n.25 (2000). The UC Regents recently voted to overturn SP-1, though Proposition 209 remains in effect. *See* Tanya Schevitz, *Affirmative-Action Ban Revoked by UC Regents*, S.F. CHRON., May 17, 2001, at A1; Rebecca Trounson & Jill Leovy, *UC Regents Vote to Rescind Ban on Affirmative Action*, L.A. TIMES, May 17, 2001, at A11.

Massachusetts,[18] and Virginia.[19] In May 2002, the Sixth Circuit, acting en banc, decided *Grutter v. Bollinger*, in which the court upheld the affirmative action program at the University of Michigan Law School.[20] The Supreme Court granted certiorari in *Grutter*, and will revisit higher education affirmative action for the first time since its landmark *Bakke* decision.[21] In addition, the Court will also review

---

15. Florida Governor Jeb Bush's November 1999 executive order replaced affirmative action in the Florida university system with the "One Florida" plan. *See Why the "One Florida" Plan Would Remove Blacks from the Best Campuses of the University of Florida*, 27 J. BLACKS HIGHER EDUC. 29, 30 (2000); Jeffrey Selingo, *What States Aren't Saying About the 'X-Percent Solution,'* CHRON. HIGHER EDUC., June 2, 2000, at 31. Governor Bush's executive order was partly a preemptive strike against an anti-affirmative action ballot initiative that Bush feared would harm his brother's presidential chances by prompting high minority voter turn-out in the November 2000 Bush-Gore election. *See* Selingo, *id.* at 32-33.

16. The voters of Washington passed Initiative 200 (I-200), a ballot initiative identical to Proposition 209, in November 1998 with a 58% majority. *See* D. Frank Vinik et al., *Affirmative Action in College Admissions: Practical Advice to Public and Private Institutions for Dealing with the Changing Landscape*, 26 J.C. & U.L. 395, 413-15 (2000). In a case involving the University of Washington Law School's affirmative action program, the Ninth Circuit recently held that racial diversity can be a compelling governmental interest that passes muster under strict scrutiny review. *See* Smith v. Univ. of Washington Law Sch., 233 F.3d 1188 (9th Cir. 2000); Sara Hebel, *U.S. Appeals Court Upholds Use of Affirmative Action in Admissions*, CHRON. HIGHER EDUC., Dec. 15, 2000, at A40; Kenneth J. Cooper, *U.S. Courts Differ on Preference; Affirmative Action Gets Mixed Verdict*, WASH. POST, Dec. 7, 2000, at A10. However, for now this is a moot point in the state of Washington because I-200 still precludes affirmative action at the University of Washington and other public institutions.

17. *See* Johnson v. Bd. of Regents of the Univ. Sys. of Georgia, 263 F.3d 1234 (11th Cir. 2001). *See also* Edward Walsh, *Court Strikes Down Georgia Admissions Policy*, WASH. POST, Aug. 28, 2001, at A5; Bill Rankin & Rebecca McCarthy, *Court Rejects UGA Effort to Enroll More Minorities*, ATLANTA J. & CONST., Aug. 28, 2001, at A1; Sara Hebel, *U. of Georgia Settles Affirmative-Action Suit*, CHRON. HIGHER EDUC., Feb. 16, 2001, at A30.

18. *See UMass Retreats From Race-Based Affirmative Action*, 27 J. BLACKS HIGHER EDUC. 12, 12 (2000); Mary Carey, *Policy or Practice?*, DAILY HAMPSHIRE GAZETTE, Mar. 26, 1999, at A1, *available at* 1999 WL 11723625; Mark Mueller, *UMass to Change Race-Based Policies*, BOSTON HERALD, Feb. 20, 1999, at 5, *available at* 1999 WL 3390642.

19. *See* Peter Schmidt, *U. of Virginia Halts Use of Scoring System That Helped Black Applicants*, CHRON. HIGHER EDUC., Oct. 22, 1999, at A42.

20. *See* Grutter v. Bollinger, 137 F. Supp. 2d 821, 825 (E.D. Mich. 2001), *rev'd en banc* 288 F.3d 732 (6th Cir. 2002). *See also* Peter Schmidt, *Appeals Court's Decision Upholding Affirmative Action in Michigan Law School Case Doesn't End Debate*, CHRON. HIGHER EDUC., May 15, 2002; Jacques Steinberg, *Court Says Law School May Consider Race in Admissions*, N.Y. TIMES, May 15, 2002, at A1.

21. *See* Grutter v. Bollinger, *cert. granted*, -- S.Ct.-- (Dec. 2, 2002), *available at* 2002 WL 1967853; *see also* Peter Schmidt, *U.S. Supreme Court Agrees to Hear 2 Affirmative Action Cases from Michigan*, CHRON. HIGHER EDUC., Dec. 2, 2002.

In *Regents of the Univ. of California v. Bakke*, 438 U.S. 265 (1978), the Court struck down the affirmative action program at the UC Davis Medical School, although it upheld the use of race as a plus factor in admission decisions. For background on the *Bakke*

Gratz v. Bollinger, a challenge to the undergraduate affirmative action program at the University of Michigan, that had yet to be decided by the Sixth Circuit.[22]

The struggle over the future of affirmative action is closely linked to the debate about how to define fairness in the meritocracy,[23] with its current emphasis on standardized tests. The *Gratz* and *Grutter* cases highlight how divergent views of standardized testing inform the opposing efforts to dismantle or defend affirmative action. In *Gratz* and *Grutter*, the principal evidence of "reverse discrimination" presented by the Center for Individual Rights (CIR) on behalf of White plaintiffs consisted of comparisons, by racial/ethnic group, of the different admission odds for applicants with similar test scores and grade point averages.[24] Thus, CIR litigation theory assumes that scores on the SAT and LSAT are a fair and adequate basis for determining who should be entitled to admission at selective colleges and universities. Given the centrality of test scores to the evidence proffered by CIR in *Gratz* and *Grutter*, and other efforts by conservative think tanks to posit SAT score differences as "incontrovertible evidence of racial preferences,"[25] affirmative action opponents are treating standardized test scores as dispositive criteria

---

case, see JOEL DREYFUSS & CHARLES LAWRENCE III, THE BAKKE CASE: THE POLITICS OF INEQUALITY (1979).

22. *See* Gratz v. Bollinger, *cert. granted*, --S.Ct.-- (Dec. 2, 2002), *available at* 2002 WL 31246645.

23. The term "meritocracy" is an invention of British Labour Party policy analyst Michael Young. Young first used this term derisively in his wicked dystopian satire. *See* MICHAEL YOUNG, THE RISE OF THE MERITOCRACY: 1870-1933: AN ESSAY ON EDUCATION AND EQUALITY (1958). For background on Young and meritocracy, see NICHOLAS LEMANN, THE BIG TEST 115-19 (1999); Nicholas Lemann, *The SAT Meritocracy: Is It Based on Real Merit?*, WASH. MONTHLY, Sept. 1997, at 32.

24. *See* William C. Kidder, *Affirmative Action in Higher Education: Recent Developments in Litigation, Admissions and Diversity Research*, 12 LA RAZA L.J. 173, 177 (2001) (summarizing the standard testing evidence presented at trial by CIR in Grutter v. Bollinger); Expert Report of Dr. Kinley Larntz, Grutter v. Bollinger, 188 F.3d 394 (6th Cir. 1999), *reprinted in* 5 MICH J. RACE & L. 463 (1999). *See also* Jodi S. Cohen, *Witness: Odds Lean to U-M Minorities*, DETROIT NEWS, Jan. 18, 2001 (summarizing Larntz's trial testimony). In *Grutter*, Larntz's testimony was found by District Court Judge Friedman to be "mathematically irrefutable proof that race is indeed an enormously important factor." Grutter v. Bollinger, 137 F. Supp. 2d 821, 841 (E.D. Mich. 2001).

25. We are referring to a series of reports on college admissions sponsored by the Center for Equal Opportunity, which is headed by Linda Chavez. *See, e.g.*, ROBERT LERNER & ANTHEA K. NAGAI, CENTER FOR EQUAL OPPORTUNITY, PERVASIVE PREFERENCES: RACIAL AND ETHNIC DISCRIMINATION IN UNDERGRADUATE ADMISSIONS ACROSS THE NATION (2001); Peter Schmidt, *Most Colleges Use Racial Preferences in Admissions, Foe of Affirmative Action Finds*, CHRON. HIGHER EDUC., Mar. 2, 2001, at A22; Douglas Lederman, *Study Documents Gaps Between White and Minority Students at Colorado Colleges*, CHRON. HIGHER EDUC., Nov. 7, 1997, at A37.

for assessing claims under the Equal Protection Clause.

In contrast, the student intervenors in *Grutter*[26] directly challenged CIR's presumption that affirmative action necessarily amounts to a preference for "lesser qualified" students of color by presenting evidence that standardized tests are racially biased.[27] The intervenors argued that affirmative action is justified in part to counterbalance the ways that tests like the LSAT and SAT taint the admissions process with racial unfairness.[28] In *Grutter,* four Sixth Circuit judges concurred that the LSAT and SAT are not race-neutral criteria for admissions. Judge Clay, joined by Judges Moore, Cole, and Daughtrey, opined that the LSAT and SAT are not race-neutral criteria for admissions and that criticism of standardized testing supports the University of Michigan Law School's consideration of race and ethnicity.[29]

Faced with the possible prohibition of using race-conscious admissions process, several states adopted "Percent Plans" that admit students based upon high school rank, without reference to SAT scores.[30] Among these are the "Ten Percent Plan" in Texas,[31] the "One

---

26. The intervenors in both *Gratz* and *Grutter* appealed separate District Court rulings prohibiting them from intervening as defendants. *See* Grutter v. Bollinger, 188 F.3d 394 (6th Cir. 1999) (consolidated cases). The Sixth Circuit overruled the two lower court rulings because it was persuaded by the intervenors' argument that the "University is unlikely to present evidence of past discrimination by the University itself or of the disparate impact of some current admissions criteria, and that these may be important and relevant factors in determining the legality of a race-conscious admissions policy." *Id.* at 401.

27. The intervenors' expert witnesses on the issue of the racial/ethnic bias on the LSAT and SAT included Martin Shapiro, Jay Rosner, David M. White, and Eugene Garcia. These four expert reports are reprinted in 12 LA RAZA L.J. 373, 377, 387, 399 (2001). *See also* Jodi S. Cohen, *Testimony Claims Law Testing Bias: Executive for Test Firm Says Questions Favor Wealthy White Males*, DETROIT NEWS, Jan. 25, 2001.

28. *See* Miranda Massie, *A Student Voice and a Student Struggle: The Intervention in the University of Michigan Law School Case,* 12 LA RAZA L.J. 231, 233 (2001) (Massie, the lead counsel for the *Grutter* intervenors argues, "[w]e engaged in a systematic critique of the manner in which racism and unearned white privilege continue to structure every aspect of educational experience in the US—and in particular, unavoidably mar the use of allegedly meritocratic criteria like LSAT scores and grades."); Defendant-Intervenors Final Reply Brief in Grutter v. Bollinger, Case No. 01-1516 (6th Cir.) July 26, 2001, at 22-26; Jodi S. Cohen, *Minorities Set to Testify at U-M Trial, Students Say Criteria Used for Law School Entry Discriminate*, DETROIT NEWS, Dec. 24, 2000, *available at* 2000 WL 30259961.

29. *See* Grutter, 288 F.3d at 769-71.

30. For analysis of percentage plans, see generally Michelle Adams, *Isn't it Ironic? The Central Paradox at the Heart of "Percentage Plans,"* 62 OHIO ST. L.J. 1729 (2001) (criticizing percentage plans because they can only succeed in preserving racial diversity in higher education if K-12 education remains racially segregated); U.S. COMM'N ON CIVIL RIGHTS, BEYOND PERCENTAGE PLANS: THE CHALLENGE OF EQUAL OPPORTUNITY IN HIGHER EDUCATION (Draft Report November 2002), *available at* http://www.usccr.gov/ (go to recent briefings and papers); U.S. COMM'N ON CIVIL RIGHTS, TOWARD AN

Florida Plan,"[32] and the "Four Percent Plan,"[33] and "12.5 Percent Provisional Admission Plan" at the University of California (UC).[34] UC President Richard Atkinson recently recommended discontinuing the use of the SAT I in favor of some other test more closely related to high school curriculum.[35] In addition, the UC Latino Eligibility Taskforce previously recommended abandoning the SAT.[36] Seen by many as an effort to dissuade the UC system, its largest customer, from abandoning the SAT, ETS recently announced a new Writing section

---

UNDERSTANDING OF PERCENTAGE PLANS IN HIGHER EDUCATION: ARE THEY EFFECTIVE SUBSTITUTES FOR AFFIRMATIVE ACTION? (April 2000), *available at* http://www.usccr.gov/ (go to publications; commission reports); Mary Francis Berry, *How Percentage Plans Keep Minority Students Out of College*, CHRON. HIGHER EDUC., Aug. 4, 2000, at A48; Jeffrey Selingo, *What States Aren't Saying About the 'X-Percent Solution,'* CHRON. HIGHER EDUC., June 3, 2000, at 31.

31. The Texas Legislature approved the "Ten Percent Plan" soon after the 1996 *Hopwood* ruling. This plan allows applicants in the top ten percent of their class to be admitted to any of the public universities in the Texas system, including selective institutions like UT-Austin and Texas A&M. For background see Danielle Holley & Delia Spencer, Note, *The Texas Ten Percent Plan*, 34 HARV. C.R.-C.L. L. REV. 245 (1999); William E. Forbath & Gerald Torres, *Merit and Diversity after Hopwood*, 10 STAN. L. & POL'Y REV. (1999); Susanna Finnell, *The* Hopwood *Chill: How the Court Derailed Diversity Efforts at Texas A&M*, *in* CHILLING ADMISSIONS 71 (Gary Orfield & Edward Miller eds., 1998); David Orenlicher, *Affirmative Action and Texas' Ten Percent Solution: Improving Diversity and Quality*, 74 NOTRE DAME L. REV. 181 (1998). We analyze the Texas Ten Percent Plan in the context of disparate impact litigation *infra* Part IV.D.iii.

32. *See generally Why the 'One Florida' Plan Would Remove Blacks from the Best Campuses of the University of Florida*, *supra* note 16.

33. The UC Regents approved the "Four Percent Plan" in March 1999. For background see Pamela Burdman, *UC Regents Rethinking Use of SAT—Newly Approved 4% Admissions Policy May Still Need Tweaking*, S.F. CHRON., Mar. 20, 1999, at A22; V. Dion Haynes, *U of California Alters Its Policy on Admissions—Change Aims to Increase Number of Minority Students*, CHICAGO TRIB., Mar. 20, 1999, *available at* 1999 WL 2855179. Likewise, in November 2001, the UC Regents approved a system-wide admissions policy that places more emphasis on special talents, overcoming adversity, and extra-curricular activities. *See* Tanya Schevitz, *UC Regents Set to Alter Admissions*, S.F. CHRON., Nov. 15, 2001, at A1.

34. In July of 2001, the UC Regents approved a type of 12.5% provisional admission plan. Under this plan, students in the top 12.5% of their high school who were not initially admitted to a UC campus can still be admitted as junior transfers (without having to reapply) if they completed two years of community college and met a certain GPA requirement specified by the UC campus. *See* Tanya Schevitz, *UC Widens Chance of Gaining Admission*, S.F. CHRON., July 20, 2001, at A1. There is no assurance that applicants under this plan can secure a spot at Berkeley and UCLA, the most selective UC campuses. *See id.*

35. *See, e.g.*, Diana Jean Schemo, *Head of U. of California Seeks to End SAT Use in Admissions*, N.Y. TIMES, Feb. 17, 2001, at A1; Kenneth R. Weiss, *SAT May Be Dropped as UC Entrance Exam*, L.A. TIMES, Feb. 17, 2001, at A1; John Cloud, *Should SATs Matter?*, TIME, Mar. 4, 2001, at 41; *see also* Selingo, *supra* note 16, at A21.

36. *See* UNIV. OF CAL. LATINO ELIGIBILITY TASKFORCE, LATINO STUDENT ELIGIBILITY AND PARTICIPATION IN THE UNIVERSITY OF CALIFORNIA: YA BASTA!, REPORT NO. 5, at 19 (1997); Z. Byron Wolf, *Task Force Urges Regents to Drop SAT Requirement*, DAILY CALIFORNIAN, Sept. 19, 1997, at 1.

and revised the Verbal section to place greater emphasis on reading comprehension and sentence completion.[37] As we will argue, however, these attempts do not mitigate the problem of disparate impact.[38]

B. *Does the SAT Accentuate or Reflect Racial and Ethnic Differences?*

A core issue underlying both "Percent Plans" and the *Grutter* and *Gratz* cases is whether standardized tests are a neutral reflection of racial and ethnic differences in educational attainment. The positions taken by many scholars and policymakers in response to this question do not correspond with their stances generally in the affirmative action debate. Rather, as will be demonstrated, a powerful conventional wisdom bridges ideological fault lines, and it is precisely this accepted wisdom that we wish to critically investigate in this empirical study.

Several "non-profit" corporations develop and market the major university undergraduate, graduate, and professional admissions tests used in American higher education, including ETS (for the SAT, GRE, and GMAT), the College Board (for the SAT), the Law School Admission Council (for the LSAT), and the American Association of Medical Colleges (for the MCAT). These organizations generally adopt liberal positions on major educational policy issues, including support for race-conscious affirmative action in higher education admissions.[39]

Former president of the College Board Donald Stewart vigorously argued against the UC Latino Eligibility Taskforce recommendation to

---

37. *See* Tamar Lewin, *College Board Announces an Overhaul for the SAT*, N.Y. TIMES, June 28, 2002 (detailing the planned changes for 2005); Tanya Schetitz, *UC's Criticisms Spur Proposal to Revise SAT Tests,* S.F. CHRON., June 18, 2002, at A4; Elizabeth Farrell, *College Board Considers Major Changes to SAT*, CHRON. HIGHER EDUC., Mar. 25, 2002 (quoting Harvard Professor Howard Gardner about UC); Eric Hoover, *SAT is Set for an Overhaul, But Questions Linger About the Test*, CHRON. HIGHER EDUC., May 31, 2002, at A35 (quoting Bob Schaeffer of FairTest about UC); Jeffrey Selingo, *U. of California Faculty Wants to Drop SAT by 2006*, CHRON. HIGHER EDUC., April 5, 2002, at A20 (reporting that the UC Regents would likely vote in July 2002 on a recommendation to drop the SAT in favor of subject exams and a writing test).

38. *See* discussion *infra* Parts III and IV.

39. *See, e.g.,* Brief of Amicus Curiae, Law School Admission Council, Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265 (1978), *reprinted in* ALLAN BAKKE VERSUS REGENTS OF THE UNIVERSITY OF CALIFORNIA: THE SUPREME COURT OF THE UNITED STATES, VOLUME IV 143 (Alfred A. Slocum ed., 1978); LAW SCHOOL ADMISSION COUNCIL, PRESERVING ACCESS AND DIVERSITY IN LAW SCHOOL ADMISSIONS – AN UPDATE (1998).

eliminate the SAT:

> It is unfortunate, as the new millennium approaches, that race, ethnic background, or family income can still limit students' educational future. Getting rid of the SAT or any other standard is not going to change that fundamental fact. Instead of smashing the thermometer, why not address the conditions that are causing the fever?[40]

Similarly, UC Santa Barbara Professor Rebecca Zwick, who spent many years as a researcher at ETS, argued that racial and ethnic gaps on the SAT are substantially equivalent to gaps in high school grades:

> Because the pattern of ethnic group differences in average high school GPA is usually similar to the pattern of average admissions test scores, an admissions policy that excludes tests but continues to include high school grades is unlikely to produce dramatic change. . . . The indisputable fact is that both high school grades and scores on admissions tests are reflections of the same education system, with all its flaws and inequities.[41]

As with testing industry insiders, a range of conservative scholars defend the SAT and other standardized tests as neutral measures of differences in educational achievement. For example, Jennifer Braceras, recently appointed by President Bush to the U.S. Commission on Civil Rights and author of a recent article defending standardized testing, concludes:

> [T]he achievement gap between black and Latino students, on the one hand, and their white peers, on the other hand, has been fount to be present across tests and across assessment devices. Thus, data from the national Assessment of Educational Progress (NAEP), the National Educational Longitudinal Survey (NELS), and the SAT all

40. Donald M. Stewart, *Why Hispanic Students Need to Take the SAT*, CHRON. HIGHER EDUC., Jan. 30, 1998, at A48. *See also* June Kronholz, *As States End Racial Preferences, Pressure Rises To Drop SAT to Maintain Minority Enrollment*, WALL ST. J., Feb. 12, 1998, at A24 (noting that the College Board rebuts the UC Latino Eligibility recommendation by arguing that eliminating the SAT would cause the White and Asian eligible pools to increase even more).

41. Rebecca Zwick, *Eliminating Standardized Tests in College Admissions: The New Affirmative Action?*, 81 PHI DELTA KAPPAN 320, 323 (1999). *See also id.* at 324 ("[B]oth test scores and high school grades are reflections of the very same disparities in educational opportunity. Eliminating standardized tests and relying more heavily on high school achievement in admissions decisions simply cannot result in a dramatic change in the ethnic diversity of the student body.").

> confirm the results of state educational assessments: African-American and Latino students lag behind their peers from other ethnic groups at every educational level. And it is not just standardized test scores that reveal this learning deficit. Grade point averages, graduation rates, and class rankings of students across the country are, regrettably, also consistent with this pattern, indicating that claims of bias are, at best, exaggerated.[42]

Similarly, Stephan and Abigail Thernstrom, influential opponents of affirmative action, reviewed evidence on class rank, grade point averages, and course selection, and concluded that the SAT gap is no larger than the gap on other measures of achievement.[43] Even more illustrative of the fact that the aforementioned conventional wisdom makes for strange bedfellows, Arthur Jensen[44] and Linda Gottfredson,[45]

---

42. Jennifer C. Braceras, *Killing the Messenger: The Misuse of Disparate Impact Theory to Challenge High-Stakes Educational Tests*, 55 VAND. L. REV. 1111, 1174-76 (2002).

43. *See, e.g.*, STEPHAN THERNSTROM & ABIGAIL THERNSTROM, AMERICA IN BLACK AND WHITE: ONE NATION, INDIVISIBLE (1997). The Thernstroms argue:

> When they heap scorn on "mere tests," defenders of affirmative action pick an easy target, and deflect attention away from a wealth of evidence demonstrating that the racial gap in other measures of academic achievement and preparation is just as large as the gap in SAT scores. . . . So far, at least, critics of tests have been unable to demonstrate that any other measure of academic preparation and achievement yields a significantly different result.

Id. at 402-03. For a critique of the conclusions the Thernstroms draw from this SAT data, see Stephen R. Shalom, *Dubious Data: The Thernstroms on Race in America*, 1 RACE & SOC'Y 125, 132-33 (1998).

44. *See* Arthur R. Jensen, *Testing: The Dilemma of Group Differences,* 6 PSYCHOL., PUB. POL'Y, & L. 121, 123 (2000) ("Nevertheless, because GPA and test scores are substantially correlated, the sole use of GPA for selection usually results in a highly similar ranking of applicants, and strict top-down selection still has almost as much adverse impact as test scores or even test scores and GPA combined."). Jensen is best known for his infamous article arguing against headstart and other social programs on the ground that IQ is largely hereditary. *See, e.g.*, Arthur R. Jensen, *How Much Can We Boost I.Q. and Scholastic Achievement?*, 38 HARV. EDUC. REV. 1 (1969); ARTHUR R. JENSEN, BIAS IN MENTAL TESTING (1980). For discussion and critique of Jensen's claims about race and IQ, see Marcus W. Feldman, *Expert Reports on Behalf of Student Intervenors: The Meaning of Race: Genes, Environments, and Affirmative Action* (expert report submitted on behalf of intervening defendants (student intervenors), Grutter v. Bollinger, 137 F. Supp. 2d 821 (E.D. Mich. 2001)(No. 97-75928)), *reprinted in* 12 LA RAZA L.J. 365 (2001); ARTHUR JENSEN: CONSENSUS AND CONTROVERSY (Sohan Modgil & Celia Modgil eds., 1987); WILLIAM H. TUCKER, THE SCIENCE AND POLITICS OF RACIAL RESEARCH (1994); Richard Delgado et al., *Can Science Be Inopportune? Constitutional Validity of Governmental Restrictions on Race-IQ Research,* 31 UCLA L. REV. 128, 136-41 (1983); Anne L. Hafner & David M. White, *Bias in Mental Research*, 51 HARV. EDUC. REV. 577 (1981).

45. *See* Linda S. Gottfredson, *Skills Gaps, Not Tests, Make Racial Proportionality Impossible,* 6 PSYCHOL., PUB. POL'Y, & L. 129, 141 (2000) (arguing that the test score gap is a neutral reflection of differences in job performance skills and concluding that "[t]ests are not the problem; banishing them is no solution. Skills gaps are the major remaining barrier to racial equality in education and employment, and therein lies the only enduring

both unabashed eugenics scholars, make arguments about the neutrality of testing that are nearly identical to those of Zwick and Stewart, respectively.

The position that the SAT, like other indicators, fairly and accurately reflects group differences in educational attainment is inconsistent with the available evidence. For example, in the last few years it was about equally as difficult for White college-bound seniors to obtain either a 600+ Verbal or 600+ Math score on the SAT as it was for them to rank in the top 10% of their high school class.[46] In contrast, it was considerably more difficult for Black and Chicano seniors to score over 600 on a section of the SAT than to rank in the top 10% of their high school class. Based on current national performance levels, even if there were equal numbers of African Americans and Whites applying to college, there still would be 4.2 times as many White as Black applicants with 600+ on the Verbal section and 5.4 times as many on the Math section.[47] The ratio is slightly lower for Chicano applicants: 3.1 White students to each Chicano student scoring 600+ on the SAT Verbal, and 3.0 Whites for every Chicano on the Math section.[48]

Yet, if we make the same kind of comparisons using high school grades, the results do not favor Whites so dramatically. Supposing there were equal numbers of Whites, Blacks, and Chicanos, the ratio of Whites to Blacks with grades in the top tenth of the class would be 1.9, and there would be "only" 1.4 times as many Whites as Chicanos among such "talented tenth" students.[49] Therefore, for Blacks and

---

solution.").

46. *See generally* College Board, 2001 Verbal and Math Percentile Ranks by Gender and Ethnic Groups, *available at* http://www.collegeboard.org/prod_downloads/about/news_info/cbsenior/yr2002/pdf/threeC.pdf (last visited Nov. 7, 2002) (reporting that among White SAT test-takers 25% had 600+ Verbal scores and 28% had 600+ Math scores) [hereinafter College Board]; Wayne J. Camara & Amy Elizabeth Schmidt, *Group Differences in Standardized Testing and Social Stratification* (1999), COLLEGE BOARD REPORT NO. 99-5, at 5 tbl.5 (reporting high school grades for 1997 college bound seniors).

47. *See* College Board, *supra* note 47.

48. *See id.*

49. *See* Camara & Schmidt, *supra* note 47. Unfortunately, Camara and Schmidt report aggregated results for all Latino (Hispanic) students combined, and do not separately report high school grades for Chicanos. In contrast, the College Board table of SAT percentile ranks separately reports various Latino subgroups, but does not report aggregate results for all Latinos combined. While this reporting difference introduces a bit of imprecision to our comparisons, it is not likely to be substantial, since SAT data suggest that other Latinos, including those from Puerto Rico, South America, and Central America, perform similar to

Chicanos applying to college, the disparate impact of requiring a 600+ on a section of the SAT is roughly twice as severe as the adverse impact of requiring graduation in the top 10% of the class. Likewise, while an equivalent proportion of White college-bound seniors obtained either an "A" average in high school or a 550+ SAT section score, for Blacks and Chicanos aspiring to go to college, a 550+ on either the SAT Verbal or Math section had almost double the impact of an "A" average.[50] This analysis is consistent with earlier representative studies documenting the adverse impact of the SAT vis-à-vis high school grades.[51]

Commissioner Braceras' argument—that standardized tests are not biased because gaps in achievement are also present in other measures[52]—is artfully imprecise and it misses the point. Few would argue that there are *no* disparities in educational measures, for what else could be expected given America's history of unequal educational opportunities? However, it hardly follows that merely because racial/ethnic educational gaps exist in grades and class rank that standardized tests are not biased. Rather, the crucial questions raised in this article are, given the consistent finding that the *magnitude* of racial/ethnic disparities in SAT scores surpasses that of other measures, why might this occur, and what are the legal and social policy implications?

II. METHODOLOGY AND RESULTS

A. *Data Samples of SAT Questions*

Representatives of the College Board and ETS often proclaim that the SAT is the single most studied test in the world. Although they are purportedly willing to provide outside researchers information about

---

Chicanos on the SAT. *See id.*

50. *See* Camara & Schmidt, *supra* note 47, at 5 tbl.5; College Board, *supra* note 47.

51. *See, e.g.*, JAMES CROUSE & DALE TRUSHEIM, THE CASE AGAINST THE SAT 92, 94 (1988) (reporting national SAT and high school rank data for the 1984 cohort of college-bound seniors); Shalom, *supra* note 44, at 132 (reporting on the SAT's greater adverse impact compared to high school grades and other measures for the 1995 cohort of college-bound seniors); William T. Dickens & Thomas J. Kane, *Racial Test Score Differences as Evidence of Reverse Discrimination: Less than Meets the Eye*, 38 INDUS. REL. 331, 338 (1999) (reporting that Black-White SAT differences are 0.30 SDs (standard deviations) greater than high school grades using a nationally representative 1982 sample from the High School and Beyond database).

52. Braceras, *supra* note 43, at 1174-76.

the SAT,[53] much of the relevant data is difficult to access. This article analyzes hard-to-access data in a fresh, original way. [54]

Our database was generated by ETS, and consists of a nationally representative sample of 100,000 test-takers who took the October 1998 SAT. The SAT currently consists of 138 test questions in its six scored sections: sixty math items and seventy-eight verbal items. Out of 138 items, 128 are multiple-choice. The remaining ten math items are called "grid-ins," and require the student to generate an answer rather than choose one from a set of four or five provided in the test booklet. After the student generates an answer, the student must then "grid-in," or bubble in his or her answer on a scantron sheet.

In addition, each SAT test-taker also answers questions from one of the many unscored experimental sections, which may include math or verbal items. The experimental section pretests new questions, and generates statistical data used to determine whether the questions should later appear as items on scored sections of future SATs. For proprietary reasons, the College Board and ETS have resisted requests for performance data on these experimental sections, even when they are more than a few years old. Consequently, our database does not permit a detailed analysis of experimental questions. However, this article does make reference to a few experimental questions that are publicly available.

To address whether the October 1998 SAT was typical with respect to racial/ethnic group performance on Math and Verbal items, we also analyzed a database with 580 questions taken from four SATs administered during 1988 and 1989. This second supplemental database includes approximately 209,000 test-takers from New York State.[55] While New York is not representative of the country's overall demographics, for our purposes it was sufficient that it included substantial proportions of African Americans (8.8%) and Latinos

53. *See* EDUCATIONAL TESTING SERVICE, ETS STANDARDS FOR QUALITY AND FAIRNESS 25 (2000) (ETS Standard 5.7 states: "Give non-ETS researchers reasonable access to ETS-controlled, nonproprietary information, if the privacy of individuals and organizations, and ETS's contractual obligations, can be met.").

54. We were initially told by several researchers and testing critics that such data was not available. Finally, Wayne Camara of the College Board put us in touch personnel at ETS, and after a series of correspondences we were able to arrange to obtain our 1998 data set for a $500 fee.

55. We thank Professor Martin Shapiro for sharing this data with us (spreadsheets on file with the authors) [hereinafter New York SAT Data].

(5.7%) in the data set.[56] Because the 1988-1989 database is older than our current sample and makes for a less representative population, this data sample was used to confirm broad conclusions about disparate impact. New York's unique "Truth in Testing" law compelled ETS to disclose these data.[57]

Our inquiry is quite straightforward: for each of the 138 items in the scored sections of the October 1998 SAT, what were the percentages of White, African American, and Chicano test-takers answering the question correctly? The impact of each question is defined as the difference between the correct responses by Whites and these racial minorities. For example, if 50% of Whites, 35% of Chicanos, and 30% of African Americans correctly answer a particular SAT Verbal question, that item has a Black-White impact of 20% and a Chicano-White impact of 15%. Part III of this article will establish that this definition of item impact is widely used by both ETS researchers and testing critics. Item impact is often associated with the *Golden Rule* procedures that emerged from a settlement between ETS and plaintiffs who sued over discrimination in the test construction process on one of ETS's standardized licensing exams.[58]

In adopting this definition of impact, it is not necessary to (and we do not) assume that all racial and ethnic differences in performance on SAT items are entirely a product of cultural bias on top of already existing disparities in preparation for higher education. Rather, our central empirical and policy question is one of degree: how much of the Black-White and Chicano-White SAT score gap can be reduced by the use of impact reduction techniques in the test development process, while still maintaining reasonable psychometric standards?

Charts 1-4 display our findings regarding the magnitude of Black-White and Chicano-White impact on the seventy-eight Verbal and sixty Math items on the October 1998 SAT. In the last few years, African Americans trailed Whites on the SAT by an average of about ninety-

56. Another significant difference is that the 1988-1989 New York data bunches all Latinos into a single category, whereas in our 1998 database we were able to separately assess Chicanos.

57. For a history and legal analysis of New York's Truth in Testing law, see Leslie G. Espinoza, *The LSAT: Narratives and Bias*, 1 AM. U. J. GENDER & L. 121, 123-25, 138-57 (1993).

58. *See infra* Part III.A-C.

five points on the Verbal section and 105 on the Math section, whereas Chicanos trailed Whites by approximately seventy-five and seventy points respectively.[59] Both the Math and Verbal sections are scored on a 200-800 scale, with a standard deviation of about 110 points.[60]

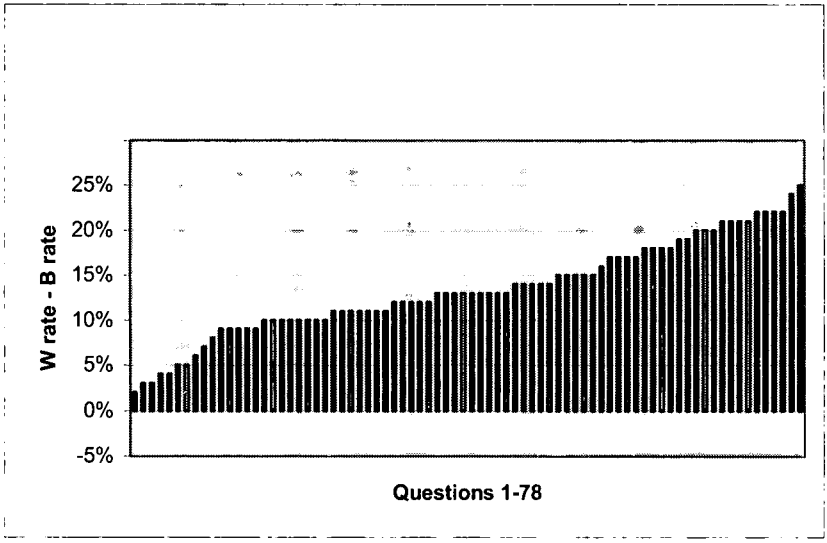
Given that a Black-White SAT average gap of approximately one standard deviation, and a Chicano-White SAT gap of about two-thirds of a standard deviation, it is hardly surprising that the percentage of Whites correctly answering each question would exceed that of African Americans and Chicanos on a substantial majority of SAT items. The consistency of the pattern, however, may be surprising: African Americans or Chicanos did not outperform Whites on any of the seventy-eight Verbal and sixty Math questions.[61] Overall, on the seventy-eight Verbal items, Whites correctly answered at an average of 59.8% and African Americans correctly answered an average of 46.4% of the items. This results in an average impact of 13.4%. Chart 1, which follows below, indicates that zero Verbal questions displayed greater African American correct response rates than White rates, and less than a tenth (7/78) of the items had differences of 5% or less. Over one-third (29/78) of the Verbal questions had Black-White differences of 15% or more, and one-sixth of the items (13/78) had gaps of 20% or more.

59. *See* College Board Press Release, *supra* note 9, at tbl.9. *See* Camara & Schmidt, *supra* note 45, at tbl.1.

60. *See* Camara & Schmidt, *supra* note 47, at tbl.1 n.2.

61. However, later we argue this pattern is not unavoidable. See *infra* Part III.

CHART 1



The pattern of disparate impact for Chicanos in Chart 2 is similar to that for Blacks, although the disparity is smaller. The overall average Chicano correct response rate was 48.7%, meaning that average Chicano-White disparate impact of the seventy-eight items was 11.1%. Out of seventy-eight items, only one question had no adverse impact on Chicanos, and just over a tenth (9/78) of the items had a disparity of 5% or less. Nearly one-fifth of the items (15/78) had gaps of 15% or greater, and three items had gaps of 20% or greater.

CHART 2

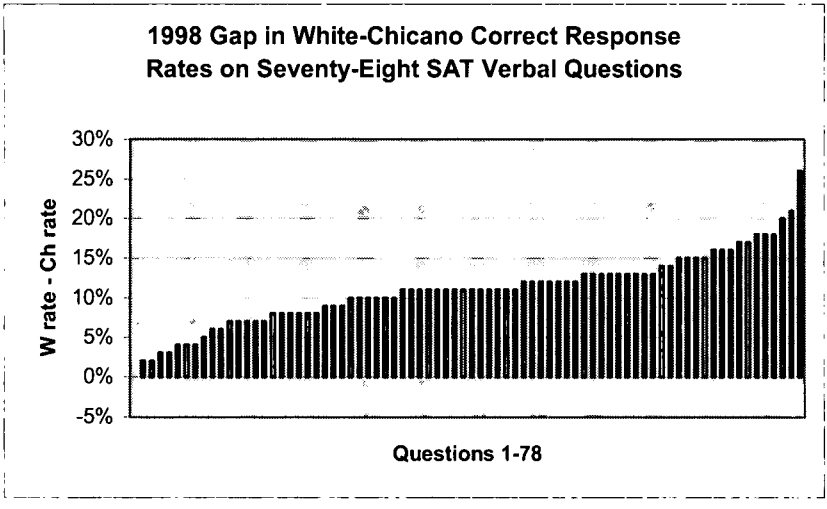
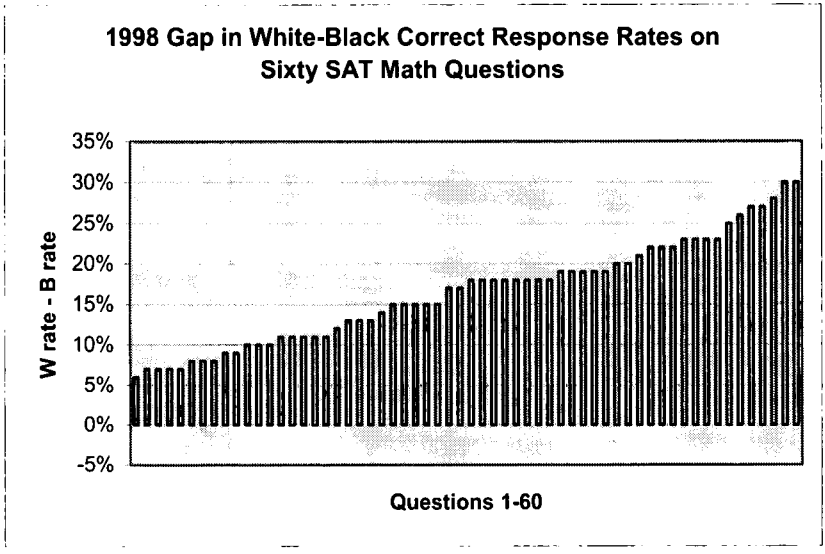


Chart 3 indicates that the Black-White disparities were somewhat larger on the Math section than on the Verbal section. Overall, the average White correct response rate was 58.4% on the sixty Math items, and the African American average correct response rate was 42.0%, for an average impact of 16.4%. One sixth (10/60) of the items had a disparate impact under 10%. Nearly three out of ten items (17/60) had a disparity of 20% or more, and two items had an impact of 30%.

CHART 3

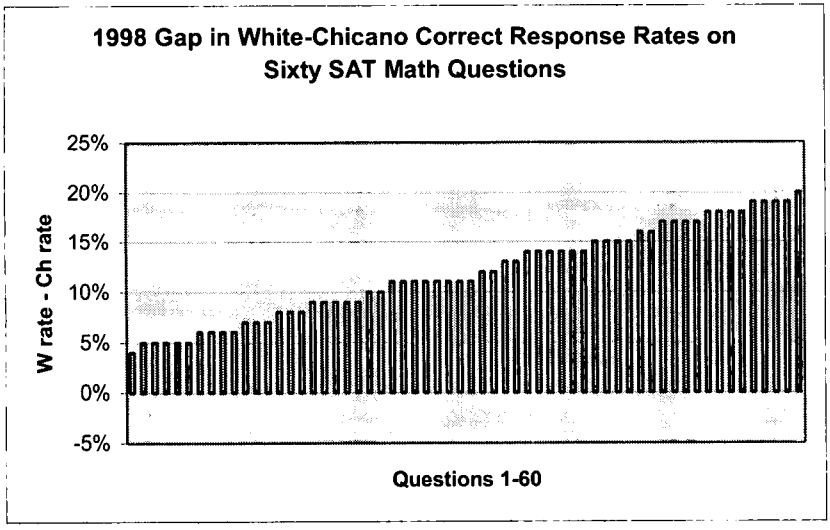


As indicated by Chart 4, the disparities for Chicanos were greater on the Math section than on the Verbal section. Overall, the average Chicano correct response rate was 46.5%, for an average disparity of 11.9% as compared to White test-takers. Interestingly however, results were not as varied (at both the low and high ends) in the Math section as on the Verbal section. This pattern may be partly attributable to bilingualism. Chicanos tend to perform relatively better on Verbal questions with vocabulary words that contain Latin root words, and they tend to perform relatively worse on Verbal items with root words that are "false cognates," which are words that appear to have Latin root words but in fact do not.[62] Only one Math item had a disparity

62. *See* Maria Pennock-Roman, *The Status of Research on the Scholastic Aptitude Test (SAT) and Hispanic Students in Postsecondary Education* 40-41 (1988), ETS RESEARCH REPORT NO. 88-36 ("For Hispanic students, bilingualism is sometimes an asset and sometimes a handicap. Items that contain English words that are true cognates of Spanish

under 5%. Nearly one-third of the items (19/60) had a disparate impact of 15% or more, and the item with the greatest disparity had a 20% gap.

CHART 4



The disparate impact of the items on the October 1998 SAT was slightly greater than that found in the 1988-89 New York SAT database. Of the 580 questions in the New York dataset, the Black-White average disparate impact was 13.2%. Specifically, for these 580 questions, Whites were more likely to answer 574 of them correctly, five items had no Black-White differences, and Blacks scored higher than Whites on one question. On the 1998 October SAT, the average

---

words in the stem and answer choices are easier, and those with false cognates are more difficult."); REBECCA ZWICK, FAIR GAME? THE USE OF STANDARDIZED ADMISSIONS TESTS IN HIGHER EDUCATION 38, 129 (2002) ("There is some evidence that Hispanic test-takers are disadvantaged by false cognates—similar words that have different meanings in the two languages."). Here is an example of an antonym problem containing a cognate in Spanish where Latinos were more likely than Whites to answer correctly. This item, presumably from the mid-to-late 1990s, was removed from the SAT by ETS at the experimental stage:

infidelity:
approval
creativity
exorbitance
loyalty (correct answer)
flightiness

Pamela Burdman, *Worth of SAT Exam Questioned*, S.F. CHRON., Nov. 11, 1997, at A1. Women also performed better than men on the same item. *See id.*

disparate impact was 14.7% for the 138 items. Moreover, Whites outperformed Blacks on all of the 138 items.

To better understand the disparate impact of each SAT item, it is helpful to examine actual SAT questions. Compare two 1998 SAT Verbal sentence completion items with similar themes: the item correctly answered by more Blacks than Whites was discarded by ETS, whereas the item that has a higher disparate impact against Blacks became part of the actual SAT. On one of the items, which was of medium difficulty, 62% of Whites and 38% of African Americans answered correctly, resulting in a large impact of 24%. The other item was pretested on the experimental section of the SAT in 1998, but it was deemed psychometrically flawed and was removed from the test. On this second item, 8% more African Americans than Whites answered correctly and 9% more women than men answered correctly.

Which Item Appeared on the SAT and Which Item was Rejected? Is Either Item (or both) Noticeably Biased?	
The actor's bearing on stage seemed _____; her movements were natural and her technique _____. a. unremitting…blasé b. fluid…tentative c. unstudied…uncontrived d. eclectic…uniform e. grandiose…controlled	The dance company rejects _____, preferring to present only _____ dances in a manner that underscores their traditional appeal. a. invention…emergent b. fidelity…long-maligned c. ceremony…ritualistic d. innovation…time-honored e. custom…ancient

The item on the left (with C as the correct answer) is the one that 8% more African Americans than Whites answered correctly. This item was omitted from the actual SAT.[63] In contrast, the item on the right (with D as the correct answer) was answered correctly by 24% more Whites than African Americans, and was included on the actual

63. This item is reported in Amy Dockser Marcus, *To Spot Bias in SAT Questions, Test Maker Tests the Test*, WALL ST. J., Aug. 4, 1999.

test.[64]

After presenting this question at several academic conferences, we found that most people cannot readily identify which item favors Whites as opposed to Blacks. As we argue at length in Part III, the facially-neutral SAT test construction will have a strong tendency to eliminate items (such as the one on the left side above) on which African Americans and Chicanos outperform Whites.

Consider another SAT Verbal item and its disparate impact. Below are two sentence completion items that are included in our data displayed in Chart 1. Whites correctly answered 59% of both items, whereas African Americans answered one of the items correctly 49% of the time, and the other 37% of the time. Can you tell which item will have a lower disparate impact of 10% and which will have a higher impact of 22%?

Which Item Will Have a Greater Black-White Disparate Impact?

The singer now performs a more _____ repertoire of songs than in the past, when he sang only traditional ballads.	Ann Wickham's marriage seemingly _____ her art because, a few years after her wedding, she began to write prolifically.
a. sentimental b. experimental c. mellow d. customary e. wary	a. quelled b. construed c. consumed d. invigorated e. sated

The item on the left (with B as the correct answer) had a Black-White disparate impact of 22%.[65] The item on the right (with D as the correct answer) had a disparity of 10%,[66] even though White test-takers found each item to be equally difficult. This article argues that a meaningful number of lower impact items can be substituted for higher impact problems without significantly compromising the psychometric

64. For verification purposes, this item is labeled VC 204 in our data set.
65. This item is labeled VC 103 in our data set.
66. This item is labeled VC 108 in our data set.

properties of the SAT.[67]

With respect to the SAT Math section, especially with items that do not include too many words or applied situations (i.e., word problems), it is difficult for many people to conceptualize how such items could be either biased against or in favor of a particular group. We argue that this difficulty is actually the point. The lack of a patently observable bias falsely implies a neutrality that does not exist. Given that educationally sound items testing similar mathematical concepts can have varying levels of disparate impact on African Americans and Chicanos, does sufficient *a priori* justification exist for preferring items that display relatively larger racial/ethnic disparities? We argue that the legitimacy of such a policy is sorely lacking, yet this is precisely what ends up happening on the real SAT and other standardized tests required for higher education admissions.

For example, in the two items below, one of the questions is from a scored SAT and was answered correctly by 11% more Whites than African Americans.[68] The other item was on the experimental SAT Math section in 1998, but was not included in a scored section of the SAT.[69] This experimental item was answered correctly by 7% more African Americans than Whites. Is it easy to distinguish the item with a disparate impact of 11% favoring Whites from the item with an impact favoring African Americans by 7%?

67. In addition, sentence completion problems such as those above will become a bigger part of the SAT starting in 2005. *See* College Board, *supra* note 47.

68. *See* New York SAT Data, *supra* note 56.

69. *See id.*

Which Item Appeared on the SAT and Which Item was Rejected? Is Either Item (or both) Noticeably Biased?	
If the area of a square is $4x^2$, what is the length of a side? a. x b. 2x c. 4x d. x^2 e. $2x^2$	If $\sqrt{2x}$ is an integer, which one of the following must also be an integer? a. $\sqrt{x}$ b. x c. 4x d. x^2 e. $2x^2$

The item on the right side (with C as the correct answer) was answered correctly by a greater percentage of African American test-takers than Whites.[70] The item on the left side (with B as the correct answer) was answered correctly by a higher percentage of Whites.[71] Would it shortchange America's high school seniors if items like that on the right appeared on the scored SAT in addition to or instead of items like that on the left?[72] While the content of both items is ostensibly neutral, can it be said that the SAT is truly unbiased if, time and time again, the test construction process tends to prefer (for statistical reasons) items like the one on the left (that favors Whites), and rejects items like the one on the right (that favors African Americans)?

III. EDUCATIONAL ANALYSIS

A. *The Devilish Details of Disparate Impact*

At the outset, we wish to make clear that neither our results nor other evidence suggests that ETS *intends* to construct the SAT and

70. *See* Marcus, *supra* note 64.

71. This item is from our New York SAT data. It was item number 14 on the second Math section of the November 1988 SAT. *See* New York SAT Data, *supra* note 56.

72. For clarification, we do not suggest that these two items specifically, which test different concepts yet have similar answer options, should be swapped. Here we remind readers that the unavailability of data on experimental questions constrains our ability to present ideally matched comparison items.

other standardized tests to ensure African Americans and Chicanos score lower than Whites. Indeed, it appears that a large majority of ETS staffers believe strongly in increased educational access for students of color and support affirmative action. We do not doubt that those involved in the development of the SAT strive for scrupulous fairness; nor do we contest that blatantly culturally biased questions (such as those rewarding familiarity with regattas, pirouettes, etc.)[73] are by far the exception and not the rule. Good intentions aside, facially neutral test construction has, for purely statistical reasons independent of discriminatory animus, the ultimate effect of contributing to—even guaranteeing—the lower performance of African Americans and Chicanos on the SAT. Obviously, our counter-intuitive and rather startling claim requires explanation.

Before any item is included in a scored section of the SAT, it must first pass through a rigorous, multi-step test construction process. The psychometricians who develop norm-referenced standardized tests such as the SAT generally adhere to two primary requirements when selecting items for the final version of the test: (1) items must be reliable, meaning that each item is internally consistent with the other items on the same test; and (2) items must meet particular specifications for level of difficulty (some questions are relatively easy, others are hard) so that the final version of the test will differentiate between test-takers of different ability levels.[74]

---

73. One example is this SAT question from the early 1980s:

RUNNER:MARATHON
(A) envoy:embassy
(B) martyr:massacre
(C) oarsman:regatta *the correct answer*
(D) referee:tournament
(E) horse:stable

On this question 53% of Whites but just 22% of African Americans chose answer (C). John Weiss, *The Golden Rule Bias Reduction Principle: A Practical Reform*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 23, 24. This question is frequently cited as an example of the SAT testing familiarity with White upper-middle class social norms rather than the ability to logically identify the appropriate relationship. *See id.*

74. *See* Robert L. Linn & Fritz Drasgow, *Implications of the Golden Rule Settlement for Test Construction*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 13 ("Classical item analysis techniques have traditionally emphasized two item characteristics: item difficulty (i.e., the proportion of test takers giving the correct answer to an item) and item discriminating power (i.e., the correlation between scores on a given item and total test scores).").

While test reliability is operationalized by means of mathematical models with forbidding names like "item response theory" (IRT),[75] the underlying concept is simple: a "reliable" item is one that people of "high ability" tend to answer correctly and people of "low ability" tend to answer incorrectly.[76] The requirements for reliability on the SAT, LSAT, GRE, and similar tests do not depend on an independent, external measure of ability.[77] Rather, item reliability is assessed by the correlation between performance on that item and performance on the test overall (or the entire portion of a test within a defined content domain).[78] If, after pre-testing, the correlation between an item and the larger test set drops below about 0.30, that item is typically flagged as a poor, unreliable question that can be excluded from the final version of the test (at least in its current form).[79]

To show how a seemingly neutral, innocuous process of selecting test questions creates an unnecessary adverse impact for students of color, imagine a pool of 1,000 pre-tested SAT Verbal sentence-completion questions in which White students, on average, score higher than Black and Chicano students.[80] Next, assume that among these 1,000 items, 100 items slightly favor Whites and 100 items slightly favor African Americans and Chicanos. As evidenced in the results section,[81] the direction (and the causes) of favoritism will seldom be readily apparent, even to expert sensitivity. After well-intentioned psychometricians calculate the correlations between each of the 1,000 items and total test scores, the key issue is which items will be accepted for an actual SAT and which items will be rejected?

Consistent with our empirical findings, we argue that in this

---

75. *See* Martin M. Shapiro, *Expert Reports on Behalf of Student Interventors: A Psychometric Model for Preserving Discrimination* (expert report submitted on behalf of intervening defendants (student intervenors), Grutter v. Bollinger, 137 F. Supp. 2d 821 (E.D. Mich. 2001)(No. 97-75928)), *reprinted in* 12 LA RAZA L.J. 387 (2001) [hereinafter Expert Report of Martin M. Shapiro].

76. *See* Jay Rosner, *Discrimination Is Built into Standardized Aptitude Tests*, LONG TERM VIEW, Sept.-Oct. 1993, at 14, 16.

77. *See* Expert Report of Martin M. Shapiro, *supra* note 76.

78. *See* Martin M. Shapiro et al., *Minimizing Unnecessary Racial Differences in Occupational Testing*, 23 VAL. U. L. REV. 213, 224-25 (1989); Linn & Drasgow, *supra* note 73, at 13.

79. *See* Shapiro et al., *supra* note 79, at 224-25.

80. This hypothetical is our adaptation of similar examples in Shapiro et al., *see id.* at 225-26; *see also* Expert Report of Martin M. Shapiro, *supra* note 70.

81. *See supra* Part II.

hypothetical, questions that are "biased" in favor of Whites have a fair chance of making their way onto a scored section of the SAT; ones that are "biased" against Whites have virtually no chance of appearing on a real SAT section.[82] Note that nothing is conspiratorial about our claim; it follows mathematically from the application of facially neutral tools of test construction. If Whites score higher overall on the set of 1,000 questions, then it must be true that "race-blind" item analysis will often detect robust and positive correlations for the items biased in favor of Whites, and weakly positive or even negative correlations for the items biased in favor of Blacks and Chicanos. In other words, item bias favoring Whites will tend to spuriously appear as reliable, whereas item bias favoring African Americans and Chicanos will, on balance, artificially appear as unreliable. The imposition of this White preference standard of test reliability necessarily follows, because the benchmark of reliability is simply the sum total of all biased and unbiased questions—meaning that there is a "tyranny of the majority" dilemma inherent in the way reliability is constructed.

While skeptics of our analysis may criticize it as too speculative, empirical evidence supports our claims. Rachelle Hackett and other ETS researchers studied the issue of disparate impact of test items by assembling two tests from a pool of experimental GRE items: the first was intended to minimize Black-White differences and the second was designed to *maximize* Black-White differences.[83] Hackett et al. found that the "maximum impact" test sections had item-test correlations that were just as high as the control group.[84] Equally troubling, the ETS researchers found that the maximum impact sections typically exhibited higher correlations with the operational (real) sections of the GRE than did the control sections.[85]

It is even more disconcerting that the disparate impact attributable to SAT reliability requirements is self-perpetuating.[86] The process of

82. *See* James W. Loewen, *A Sociological View of Aptitude Tests*, *in* U.S. COMM'N ON CIVIL RIGHTS, THE VALIDITY OF TESTING IN EDUCATION AND EMPLOYMENT 73, 85-86 (1993) (noting that point-biserial requirements are more likely to exclude items favoring women and minorities) (citing DAVID OWEN, NONE OF THE ABOVE: BEHIND THE MYTH OF SCHOLASTIC APTITUDE 124 (1985)).

83. *See* Rachelle Kisst Hackett et al., *Test Construction Manipulating Score Differences Between Black and White Examinees: Properties of the Resulting Tests* 31 (1987), ETS RESEARCH REPORT NO. 87-30.

84. *See id.* at 8 tbl.3.

85. *See id.* at 18 fig.10.

86. *See* Shapiro et al., *supra* note 79, at 226; Expert Report of Martin M. Shapiro, *supra* note 76.

developing new SAT questions is an ongoing feedback loop that includes writing, pre-testing, analyzing, and finally administering scored questions. Test writers, regardless of their background, are rewarded for maximizing the number of "reliable" items they construct, and minimizing the wasted time associated with developing items that will later be tossed away as "unreliable." Thus, this subtle White preference standard may become an imbedded social norm over the course of successive test administrations.[87] Such a bias tends to be obscured because Whites have historically scored higher on the SAT than African Americans and Chicanos. The entire score gap is usually attributed to differences in academic preparation, although a significant and unrecognized portion of the gap is an inevitable result of the flaw in the development process.[88]

Our second, more concrete example also sheds light on the consequences of developing standardized test items around traditional notions of reliability. Over the years, one of us (Mr. Rosner) provided pro bono legal services to students in disputes with standardized test producers. One such student, Chris Laucks, took the LSAT in 1981 when it included math problems similar to those appearing on the SAT or GRE today. On a particular geometry problem, ETS mistakenly omitted a right angle marker. With the marker, one answer would have been mathematically correct, but in the absence of the marker, a different answer was mathematically correct. Unfortunately, Laucks picked the answer he knew to be mathematically correct instead of the answer he suspected ETS wanted.

After Mr. Laucks received his LSAT score with this geometry question marked incorrect, he wrote a complaint to the Law School Admission Council (LSAC) and attached a flawless mathematical proof of his answer. Oddly enough, LSAC confirmed that Laucks was correct, but it would not give him credit for his correct answer. Strict adherence to correlation requirements accounts for LSAC's peculiar stance. In its view this item was defective, because in pre-testing, "high ability" students picked the other, incorrect answer. Thus, to give Laucks points for this question (and to penalize those picking the

87. *Cf.* Daria Roithmayr, *Barriers to Entry: A Market Lock-in Model of Discrimination*, 86 VA. L. REV. 727 (2000) (advancing an economic argument that institutional networks among law school professional organizations tend to construct merit criteria that pose significant barriers to entry for people of color, and that such discrimination becomes imbedded over time).

88. *See* Shapiro et al., *supra* note 79, at 226.

other answer) would violate a sacrosanct principle of test reliability. Accordingly, ETS wrote a letter to test-takers explaining why the item was pulled from scoring (rather than crediting Laucks and others who picked (D) as the correct answer), in which it stated:

> As was noted above, the credited response to the question was (C). Statistical results from a trial administration of the question indicated that the question, with (C) as the answer, was functioning as intended. If the question had been keyed (D) in the trial administration, the statistics would have shown that it did not function properly, and it would not have been used in the LSAT.[89]

Chris Laucks learned the hard way that traditional psychometric methods will not allow the "right" people to get the answer wrong and the "wrong" people to get the answer right, even if this is what happened in fact. Critical race theorist Richard Delgado has criticized standardized tests for their "epistemological fascism" because of the ways such tests reward particular thinking styles and punish other styles.[90] Laucks' *Alice in Wonderland* experience with the LSAT—if the highest scorers pick "A" as their answer, and it is later proven that "B" is the correct answer, then the question, and not the answer key is deemed to be defective—is certainly consistent with Delgado's criticism.

While test producers vigorously defend item-test reliability as an essential tool of sound test construction, examples such as Laucks' raise the point that the overzealous pursuit of test reliability can actually undermine the construct validity[91] of the SAT and similar

89. Letter from ETS to LSAT Test-Takers, April 30, 1981, *reprinted in* DAVID M. WHITE, THE EFFECTS OF COACHING, DEFECTIVE QUESTIONS, AND CULTURAL BIAS ON THE VALIDITY OF THE LAW SCHOOL ADMISSION TEST Appendix A (1984).

90. *See* Richard Delgado, Barrett Lecture on Constitutional Law at UC Davis Law School (Oct. 12, 2000). *See also* Richard Delgado, *Official Elitism or Institutional Self-Interest? 10 Reasons Why UC Davis Should Abandon the LSAT (and Why Other Good Law Schools Should Follow Suit),* 34 U.C. DAVIS L. REV. 593, 599 (2001) ("Standardized tests punish takers who deviate from the path the designer has in mind. This enforced orthodoxy is independent of particular items and terms that disadvantage minorities and the working class, such as regattas and tuxedos. It also punishes those who think outside the box."); Lani Guinier, *Confirmative Action,* 25 LAW & SOC. INQUIRY 565, 582 (2000) ("One can certainly begin to speculate, however, that multiple-choice, timed testing may train successful candidates not to question authority, not to look for innovative ways to solve problems, not to do sustained research or to engage in team efforts at brainstorming, but instead to try to answer questions quickly and in ways that anticipate the desires or predilections of those asking the questions.").

91. For a discussion of construct validity see, e.g., U.S. DEPT OF EDUC. OFFICE FOR CIVIL RIGHTS, THE USE OF TESTS AS PART OF HIGH-STAKES DECISION-MAKING FOR

standardized tests. For example, Professor Stuart Katz and his colleagues at the University of Georgia gave SAT Reading Comprehension questions and answers to students *without* the actual reading passages. Katz found that because of factors such as outside knowledge and test-wiseness, honors students correctly answered forty-seven of the 100 questions on average, and a broader mix of students answered thirty-eight questions correctly, whereas random guessing would result in about 20 correct responses.[92] However, SAT Reading Comprehension sections were altered in the 1990s to include fewer but longer passages with more questions, and Reading Comprehension increased from 29% of the Verbal score to 51%.[93] Despite these revisions to the SAT, Katz found that students in introductory psychology courses could still answer 36% of the new Reading Comprehension items correctly without access to the reading passages.[94] We suggest, based partly on this line of research, that reliability requirements likely play a role in undermining the construct validity of SAT Reading Comprehension. If savvy test-taking is more helpful than actually understanding the reading passages, then the SAT's construct validity is suspect, and the hypertrophied virtue of item reliability may be contributing to the degradation of construct

---

STUDENTS 25 (Dec. 2000) ("Construct validity refers to the degree to which the scores of test takers accurately reflect the constructs a test is attempting to measure."); AM. EDUC. RESEARCH ASS'N ET AL., STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING 173 (1999) (defining a construct as "the concept or the characteristic that a test is designed to measure"); Samuel Messick, *Foundations of Validity: Meaning and Consequences in Psychological Assessment*, ETS RESEARCH REPORT NO. 1, at 9 (1993) (stating that construct validity "comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables") Samuel Messick, *Validity*, *in* EDUCATIONAL MEASUREMENT, THIRD EDITION 13, 42 (Robert L. Linn ed., 1989) ("Indeed, the substantive component of construct validity entails a veritable confrontation between judged content relevance and representativeness, on the one hand, and empirical response consistency, on the other.").

92. *See* Stuart Katz, *Answering Reading Comprehension Items Without Passages on the SAT*, 1 PSYCHOL. SCI. 122, 123, 125 (1991). *See also* Chris Raymond, *Study Questions Validity of Reading-Comprehension in SAT*, CHRON. HIGHER EDUC., April 25, 1990, at A5 (describing a Katz study and response by the College Board).

93. *See* Stuart Katz, *Answering Reading Comprehension Items Without Passages on the SAT-I*, 85 PSYCHOL. REP. 1157, 1158 (1999).

94. *See id.* at 1160. For further corroboration of this line of research see Stuart Katz et al., *Answering Reading Comprehension Items Without Passages on the SAT When Items Are Quasi-Randomized*, 51 EDUC. & PSYCHOL. MEASUREMENT 747 (1991); Stuart Katz & Gary J. Lautenschlager, *Answering Reading Comprehension Questions Without Passages on the SAT-I, ACT and GRE*, EDUC. ASSESSMENT 295 (1994); Stuart Katz & Gary J. Lautenschlager, *The SAT Reading Task in Question: Reply to Freedel and Kostin*, 6 PSYCHOL. SCI. 126 (1995); Stuart Katz et al., *Answering Quasi-Randomized Reading Items Without the Passages on the SAT-I*, 93 J. EDUC. PSYCHOL. 772 (2001).

validity.

B. *Does Differential Item Functioning Eliminate or Exacerbate Item Bias?*

Our principal claim—that SAT reliability requirements can facilitate test item bias against Black and Chicano students—would be weakened if ETS and other test developers used methods that dependably rooted out biased items in the first place. Differential Item Functioning (DIF) is a statistical technique for identifying specific test items that are disproportionately more difficult for members of a race or gender group among test takers with *equivalent overall test scores*.[95] The Mantel-Haenszel statistic and "standardization" are two very similar methods, and are used by ETS, LSAC, and other test developers to measure DIF.[96]

ETS promotional materials suggest that DIF is a sound method for flagging items that can unfairly penalize minorities. Sydell Carlton of ETS states:

> Matching students according to their test scores and then examining how they did on individual test questions helps us to determine whether the test questions themselves may be creating problems for a particular group. . . . By using the DIF procedure, paired with the Test Sensitivity Review procedure, ETS helps ensure that its examinations provide a level playing field for all who take them.[97]

---

95. For an in-depth discussion of DIF techniques, see generally DIFFERENTIAL ITEM FUNCTIONING (Paul W. Holland & Howard Wainer eds., 1993).

96. *See* W. Edward Curley & Alicia P. Schmitt, *Revising SAT-Verbal Items to Eliminate Differential Item Functioning*, COLLEGE BOARD REPORT NO. 93-2, at 3-4 (1993) (reviewing these two procedures and noting that they produce highly similar results); Loewen, *supra* note 83, at 84.

97. Educational Testing Service, *What's the DIF? Helping to Ensure Test Question Fairness*, *at* http://www.ets.org/research/dif.html (last visited Dec. 31, 2001). This claim is not atypical of ETS and other test developers. *See, e.g.*, Curley & Schmitt, *supra* note 91, at 3 ("Since DIF indices take into account overall differences in ability on the construct being measured by matching the groups before comparing their performance, DIF indices identify items that might have construct-irrelevant characteristics."); Jane Faggen, *Golden Rule Revisited: Introduction*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 5, 7 ("The Mantel-Haenszel statistic helps to identify differences in performance on an item-by-item basis that may reflect potentially irrelevant characteristics in certain test questions that may be unfair to certain groups."); Richard M. Jaeger, *NCME Opposition to Proposed Golden Rule Legislation*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 21, 22 (President of the National Council on Measurement in Education's (NCME) public letter to the New York legislature in opposition to a bill adopting item bias methods similar to those we advocate in this article: "Currently accepted measures of test item bias do not rest upon average performance differences between groups. Evidence of bias requires that an item be found to perform differently *for individuals of equal ability*.").

As with claims about the SAT not having a disparate impact relative to other educational measures, ETS's public stance with respect to DIF does not withstand careful inspection. Put simply, DIF does not and cannot, as Carlton argues, "provide a level playing field." Continuing with this same metaphor, DIF techniques actually *assume* an overall level playing field, then proceed to look for an unusual pothole that might unfairly trip up one team or another, so to speak. If, for example, the playing field favors the home team by allowing them to run downhill to score a goal and forcing the away team to run uphill to score, this obvious bias would be undetected by DIF. The ETS "level playing field" argument is misleading and circular; by controlling for total test score before looking for potentially biased items, it is not possible for DIF to remove aggregate bias or lessen the overall racial and ethnic score gaps on the SAT.[98]

ETS and other researchers even argue that since DIF does not decrease racial disparities, this is further corroboration that SAT items were unbiased all along.[99] Needless to say, we find this logic unconvincing. As James Loewen aptly put it, "DIF removes the adverse impact before looking for adverse impact!"[100] A close look at the educational measurement literature reveals that several esteemed psychometricians, including many working for ETS and other test producers, acknowledge that DIF cannot identify and eliminate systematic item bias against a minority group because controlling for total test score means there is no external fairness standard.[101]

---

98. *See* Shapiro et al., *supra* note 79, at 226 ("[T]he available psychometric measures of item bias do not measure item bias *per se* but only item bias relative to overall test bias. These methodologies can only detect whether a particular item is significantly more biased or significantly less biased than the aggregate of all the test items as a whole."); Loewen, *supra* note 77, at 84 (noting that DIF does not impact group averages on a test).

99. *See* Elizabeth Burton & Nancy W. Burton, *The Effect of Item Screening on Test Scores and Test Characteristics*, *in* DIFFERENTIAL ITEM FUNCTIONING, *supra* note 96, at 321; ZWICK, *supra* note 63, at 130. *See also* John E. Hunter & Frank L. Schmidt, *Racial and Gender Bias in Ability and Achievement Tests: Resolving the Apparent Paradox*, 6 PSYCHOL. PUB. POL'Y & L. 151 (2000).

100. Loewen, *supra* note 83, at 85. Recall that our main point is that the test assembly procedures overall, rather than DIF specifically, worsens disparate impact.

101. *See, e.g.*, William H. Angoff, *Perspectives on Differential Item Functioning Methodology*, *in* DIFFERENTIAL ITEM FUNCTIONING, *supra* note 96, at 3, 17 ("For if the criterion is itself biased to some degree, then the application of a DIF analysis will certainly be flawed; further, if bias is pervasive in the criterion, then any attempt to identify bias in its component items will inevitably fail."); Lorrie Shepard et al., *Comparison of Procedures for Detecting Test-Item Bias with Both Internal and External Ability Criteria*, 6 J. EDUC. STAT. 317, 321 (1981) ("A major limitation of all of the bias detection approaches employed in the

Group test averages cannot be changed by DIF, which creates the foregone conclusion that questions biased "against" a group are counterbalanced by questions "in favor" of that group.[102] Some experts even argue that DIF can exacerbate rather than eliminate item bias against students of color because many questions favoring Whites would not stand out statistically after controlling for overall test score.[103]

C. *Can* Golden Rule *and Sound Test Development Procedures Coexist?*

Our approach to reducing test item bias on the SAT bears some resemblance to the *Golden Rule* technique for ameliorating racial item bias, so this portion of Part III addresses common criticisms of the *Golden Rule* procedures. *Golden Rule* was a 1984 settlement of a lawsuit brought by the Golden Rule Insurance Company against the ETS over alleged racial bias on the Illinois Insurance Exam.[104] The core principle underlying this settlement was that when items are selected for the final version of the test, questions in each content area having smaller Black-White differences should be preferred over questions in the same content domain with larger racial disparities.[105] This principle was operationalized by classifying all questions as either Type I or Type II items after pre-testing. Type I items were those with Black-White correct answer rate differences of 15% or less and overall

---

research to date is that they are all based on a criterion internal to the test in question. They cannot escape the circularity inherent in using total score on the test or the average item to identify individuals of equal ability and hence specify the standard of unbiasedness."); Nancy S. Cole, *Judging Test Use for Fairness*, *in* U.S. COMM'N ON CIVIL RIGHTS, THE VALIDITY OF TESTING IN EDUCATION AND EMPLOYMENT 92, 102 (1993) (Cole, former ETS President, acknowledged that DIF cannot "guarantee that there is no gender bias in the questions."); Howard Wainer, *Precision and Differential Item Functioning on a Testlet-Based Test: The 1991 Law School Admissions Test as an Example*, 8 APPLIED MEASUREMENT IN EDUC. 157, 182 (1995) (conducting an ETS-sponsored study of LSAT DIF and noting, "Because performance on the test section itself determined the stratifying variable, the overall balance (zero overall DIF) is almost tautological. That the balancing works as well as it does at all levels of examinee proficiency is not mathematically determined.").

102. *See* Gregory Camilli, *The Case Against Item Bias Detection Techniques Based on Internal Criteria, in* DIFFERENTIAL ITEM FUNCTIONING, *supra* note 96, at 397, 409 ("Holding ability constant, if one group of examinees tends to miss some items unexpectedly, it must unexpectedly answer other items correctly. In other words, items that disfavor the minority group are canceled by items that favor the minority group.").

103. *See* Loewen, *supra* note 83, at 85-86.

104. *See* Patrick Rooney, *Golden Rule on Golden Rule*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 9, 10-11 (discussing Golden Rule Ins. Co. v. Washburn, No. 419-76 (Ill. Cir. Ct. 7th Jud. Cir. Nov. 20, 1984) (consent decree)).

105. *See* Shapiro et al., *supra* note 79, at 250-52.

correct answer rates of 40% or higher.[106] Type II referred to all other items, including those with large racial disparities. Four terms of the settlement covered these two categories: (1) Type I items were to be used as long as they were available in sufficient numbers; (2) among Type I items, those with the smallest Black-White disparities were to be used first; (3) Type II items could be used when Type I items were not sufficiently available; and (4) among Type II items, those with the smallest Black-White disparities were to be used first.[107] After *Golden Rule*-type procedures were proposed in litigation over the National Teacher Exam in Alabama and then in college admission testing legislation in New York and California, ETS announced that the *Golden Rule* settlement was a "mistake."[108]

The debate over the *Golden Rule* settlement is important for our findings. As psychometrician Lloyd Bond observed, "The psychometric profession is virtually unanimous in its condemnation of the *Golden Rule* as a bad precedent, even if unintended."[109] One criticism made by psychologists and lawyers for testing corporations is that the *Golden Rule* approach will jeopardize the "blueprint" of standardized tests, which has to do with the balancing of different forms of content on the exam.[110] For instance, education attorney Michael Rebell hypothesized that on a teacher exam, if minority teaching candidates do perform relatively worse on geometry problems, then the *Golden Rule* method will distort the test by unwisely steering the test away from an educationally optimal weighting of geometry problems.[111] However, Rebell's criticism is a straw man argument, insofar as the *Golden Rule* settlement carried out the classification of test item types separately *within* each subject area.[112] Likewise, we advocate minimizing racial and ethnic performance disparities *within* each of the SAT subsections and content areas.

---

106. *See* Faggen, *supra* note 98, at 5, 6.

107. *See* Linn & Drasgow, *supra* note 75, at 13, 14.

108. *See* Faggen, *supra* note 98, at 6; Robert L. Linn, *Bias in College Admissions Measures*, *in* THE COLLEGE ADMISSIONS PROCESS: A COLLEGE BOARD COLLOQUIUM 80, 81 (1986); Gregory R. Anrig, *ETS on "Golden Rule,"* EDUC. MEASUREMENT ISSUES & PRACTICE, Fall 1987, at 24.

109. Lloyd Bond, *The Golden Rule Settlement: A Minority Perspective*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 18, 20.

110. *See* Michael A. Rebell, *Disparate Impact of Teacher Competency Testing on Minorities: Don't Blame the Test-Takers—or the Tests,* 4 YALE L. & POL'Y REV. 375, 394 (1986); S.E. Phillips, *The Golden Rule Remedy for Disparate Impact of Standardized Testing: Progress or Regress?*, 63 ED. LAW REP. 383, 412-13 (1990).

111. *See* Rebell, *supra* note 111, at 394.

112. *See* Shapiro et al., *supra* note 79, at 250 n.164.

The most common criticism of *Golden Rule* procedures is related to the "test blueprint" issue. Linn, Dragow, Rebell, and Jaeger argue that *Golden Rule* will disproportionately allow only easy items on the scored version of the test, because difficult items create the largest racial and ethnic gaps in performance.[113] This criticism lacks empirical support. In our database of 1998 SAT questions, we generally found that group differences are smaller on difficult and easy questions, and largest on questions of moderate difficulty.[114] For example, Robert Linn of the University of Colorado, one of the more outspoken critics of *Golden Rule*, states that "such an approach would have a negative effect on the reliability and validity of the resulting tests."[115] To support his argument that *Golden Rule* "tortures validity,"[116] Linn claims that only a meager proportion of SAT items would qualify as Type I items under *Golden Rule*.[117] However, Linn's comparison between the SAT and the Illinois Insurance Exam is unpersuasive. Each question on the Insurance Exam only has four options, whereas SAT (and LSAT, GRE, GMAT, etc.) items have five options. Thus, Linn compares apples to oranges when he claims that *Golden Rule*'s 40% minimum correct threshold will exacerbate rather than lessen group test score differences because it would tend to eliminate SAT items with the smallest disparities.[118] Moreover, such a critique is irrelevant to our results about the SAT, since we believe it would be unnecessary to impose a minimum cut-off for correct answer rates. In the context of the Illinois Insurance Exam, the 40% threshold was merely an attempt to ensure that selected items have a correct-rate above random guessing (25%, based on four multiple choice

113. *See* Linn, *supra* note 109, at 81; Rebell, *supra* note 111, at 394; Linn & Drasgow, *supra* note 75, at 14-15; Richard M. Jaeger, *supra* note 98, at 21, 22.

114. We note that this is partly a consequence of defining impact based on the difference in percentage correct rates. For example, a question answered correctly by 20% of Whites and 15% of Blacks appears small because 20% minus 15% equals 5%. In contrast, one could use another definition, such as the ratio of correct rates. From the latter perspective, African American performance would only be 75% of White performance, a seemingly larger discrepancy, and one that is not necessarily smaller than performance differences on typical moderate-difficulty questions. Nonetheless, we still argue that Linn, Drasgow, Rebell, and Jaeger are incorrect because these scholars reference the same definition of impact that we adopt in this article.

115. Linn, *supra* note 109, at 81.

116. *Id.*; Linn & Drasgow, *supra* note 75, at 17.

117. Linn, *supra* note 109, at 81 (claiming only twenty-five of eighty-five SAT items would be classified as Type I). While Linn raises this point in order to condemn *Golden Rule*, a skeptic of standardized testing might interpret the same data as an admission that a high proportion of SAT items are indeed tinged with racial bias.

118. *See* Linn, *supra* note 109, at 81; Linn & Drasgow, *supra* note 75, at 14-15.

questions).[119] Moreover, for the SAT, the large population of 1.3 million test-takers eases concerns about inadequate item pools.

In their enthusiasm to condemn *Golden Rule*, Linn and Drasgow advance the curious position that *Golden Rule* will: (1) corrupt test validity because it will eliminate items with the largest group disparities, which they claim tend to be the most difficult items; and (2) worsen racial disparities because elimination of proportionately more difficult items will generally remove the items with smaller racial/ethnic differences.[120] As we demonstrate below, these claims are unsubstantiated and are contrary to subsequent empirical research, much of which was conducted by ETS.

In reality, the *Golden Rule* method in fact decreased Black-White differences on the Illinois Insurance Exam.[121] Additionally, when ETS researchers applied *Golden Rule*-inspired adverse impact reduction procedures to experimental sections of the GRE, they acknowledged that racial/ethnic disparities could be lessened without compromising test integrity:

> First, such techniques *can* reduce impact . . . . Second, the resulting tests *can* be made to look parallel in form and content to conventionally constructed tests and meet their content specifications if the item pools are sufficiently large. Third, the *average* difficulty level of the resulting tests *can* be maintained without changing current test development procedures for adhering to average difficulty specifications. However, the *distribution* of item difficulties will change . . . . This may be a controllable phenomenon.[122]

Unfortunately, in the fifteen years since this ETS study was published, ETS, the College Board, LSAC, GMAC, and AAMC still have not implemented impact reduction techniques on the SAT, LSAT, GRE, GMAT, or MCAT. Martha Stocking and other ETS researchers recently revisited the issue of item bias reduction techniques on populations of women, African Americans, and Latinos.[123] Their

---

119. *See* Shapiro et al., *supra* note 79, at 251 n.166.

120. *See* Linn, *supra* note 109, at 81; Linn & Drasgow, *supra* note 75, at 14-15.

121. *See* Shapiro et al., *supra* note 79, at 254-55; John Weiss, *The Golden Rule Bias Reduction Principle: A Practical Reform*, EDUC. MEASUREMENT ISSUES & PRACTICE, Summer 1987, at 23, 25. *But see* Phillips, *supra* note 111, at 404-05 (questioning claim of Shapiro et al. that the *Golden Rule* technique reduces racial disparities).

122. Hackett et al., *supra* note 84, at 31.

123. *See generally* Martha Stocking et al., *An Empirical Investigation of Impact*

findings, consistent with our findings and earlier research, were that accounting for group differences when assembling SAT test forms *can* lessen the adverse impact of the test *without* compromising construct validity and with only minor effects on test reliability.[124]

D. *Practical Considerations*

1. *What Are the Consequences for Asian Pacific Americans and for Women?*

SAT disparate impact reduction procedures raise thorny policy questions about race, ethnicity, gender, and other categories. Which groups should be included in efforts to reduce adverse impact on standardized test questions, and what are the consequences of excluding certain groups from *Golden Rule*-style adjustments?[125] We will briefly discuss two key considerations: (1) the impact on Asian Pacific American (APA) test-takers;[126] and (2) the feasibility of simultaneously reducing adverse impact for African Americans, Latinos,[127] and women.

Because of the role that education plays in America's opportunity structure, it is particularly important to attend to interracial conflicts that may arise from our approach to impact moderation on the SAT.[128]

Moderation in Test Construction, ETS RESEARCH REPORT NO. 01-04 (2001). While the authors of this study did not disclose the particular Math and Verbal test they studied, the details of their study—including the racial, ethnic and gender gaps on the test, and the size of the populations taking each of the test forms and the size of the item pools—strongly suggest that this was a study of the SAT. *Cf. id.* at 7 tbl.1.

124. *See id.*

125. *Cf.* Paul Brest & Miranda Oshige, *Affirmative Action for Whom?*, 47 STAN. L. REV. 855 (1995) (analyzing similar questions in the context of education and employment affirmative action programs).

126. The term APA is extremely heterogenous. Unfortunately, the College Board appears not to publish annual data on the composition of APA students taking the SAT by subgroup. Data from UC Berkeley is informative on this point, but is probably not representative of national trends. *See* Mark Tanouye et al., Asian Pacific Americans at Berkeley: Visibility and Marginality 17 (2001) (unpublished report by the UC Berkeley Campus Advisory Committee for Asian American Affairs to UC Berkeley Chancellor Robert Berdahl) (In 2000, there were 9,110 APAs at Berkeley (40% of the undergraduate student body), and of this group 50% had national origins in China, 15% in Korea, 9% in India/Pakistan, 7.5% in Vietnam, 7.5% in the Philippines, 5% in Japan, and 1% in the Pacific Islands). *See id.*

127. While we looked at Chicanos specifically in our data set, the ETS research we cite to on this point covers Latinos rather than Chicanos specifically. *See infra* note 142 and accompanying text.

128. *See, e.g.*, ERIC K. YAMAMOTO, INTERRACIAL JUSTICE: CONFLICT AND RECONCILIATION IN POST-CIVIL RIGHT AMERICA (1999) (analyzing interracial conflict in

Some readers may have legitimate concerns about whether using item bias reduction techniques in order to produce a "fairer" SAT for African Americans and Latinos might unintentionally cause harm to APAs taking the SAT. APAs comprised 8% of those taking the SAT in 1991, and this grew to 10% in 2001.[129] Over the last decade APAs have scored about thirty points lower on average than Whites on the Verbal section of the SAT and about thirty-five points higher than Whites on the Math section.[130]

Stocking's most recent ETS study of impact moderation on the SAT indicates that attempts to reduce Black-White and Latino-White test score gaps will not adversely effect APAs.[131] In a sample of 5,863 APA test-takers, the four methods of moderating the Verbal section of the test resulted in an average increase in the gap favoring Whites by 0.015 standard difference units, and the six methods of moderating the Math section resulted in increasing APAs' advantage by 0.083 standard difference units.[132] In practical terms, this would translate to a net gain for APAs of approximately five points on the SAT.[133] In fact, the Verbal section which resulted in the best impact reduction for Blacks and Latinos also most effectively decreased APAs' disadvantage on the Verbal section, whereas the section that increased disparate impact for Blacks and Latinos negatively affected APAs as well.[134] Likewise, the Math section which most effectively reduced impact for African Americans and Latinos also increased APAs' advantage on the Math section vis-à-vis Whites.[135] These findings suggest no inherent conflict between impact reduction techniques and APA performance on the

education and other settings); *see also* Kevin R. Johnson, *Lawyering for Social Change: What's a Lawyer to Do?*, 5 MICH. J. RACE & L. 201 (1999).

129. *See* College Board Press Release, *supra* note 9, at 6.

130. *See id.*; *How Scores on the SAT Vary*, CHRON. HIGHER EDUC., Sept. 17, 1999; Eric Hoover, *Average Scores on the SAT and the ACT Hold Steady*, CHRON. HIGHER EDUC., Sept. 7, 2001, at A52; Leo Reisberg, *Disparities Grow in SAT Scores of Ethnic and Racial Groups,* CHRON. HIGHER EDUC., Sept. 11, 1998, at A42.

131. *See* Stocking et al., *supra* note 124, at 15 tbl.4; WARREN W. WILLINGHAM & NANCY S. COLE, GENDER AND FAIR ASSESSMENT 21-23 (1997).

132. Standard difference (D) is a statistic used to make group comparisons across different tests and populations. *See* WILLINGHAM & COLE, *supra* note 132, at 21-23.

133. This rough estimate is extrapolated from Willingham and Cole's chart listing D values on the SAT. *See id.* at 24 fig.2.2.

134. Stocking et al., *supra* note 124, at 15 tbl.4. (showing that application of the "test construction" method to Verbal Section 2 decreased the gaps by 0.18 Ds for African Americans, 0.15 Ds for Latinos, and 0.10 Ds for APAs while application of the "test selection" method to Verbal section 1 increased the gaps by 0.04 Ds for African Americans, 0.09 Ds for Latinos, and 0.13 Ds for APAs).

135. *See id.* (showing that application of the "test construction–small" method to Math Section 2 decreased the gaps by 0.12 Ds for African Americans, 0.07 Ds for Latinos, while it increased APA performance by 0.17 Ds).

SAT.

In summary, we are confident in concluding that *Golden Rule*-style impact moderation techniques would not create a barrier to opportunity for APA college applicants. Other factors—such as legacy preferences at elite private universities, SAT Verbal cut-off scores, and covert enrollment ceilings—pose far more serious threats to equal educational opportunity for APAs in the contemporary admissions environment.[136]

Standardized tests usually have a modest disparate impact on women, as it is well documented that women perform slightly less well than men (both overall and within racial/ethnic groups) on the SAT, GRE, GMAT, MCAT, and LSAT.[137] This pattern occurs despite the fact that women consistently obtain better grades than men in high school, college, and most graduate school programs.[138] Consequently, higher education standardized tests are frequently criticized for being gender biased.[139]

Fortunately, ETS research repeatedly demonstrates that it is

---

136. *See, e.g.*, DANA Y. TAKAGI, THE RETREAT FROM RACE: ASIAN-AMERICAN ADMISSIONS AND RACIAL POLITICS 34, 62-70, 96-98 (1992); *see* Grace W. Tsuang, Note, *Assuring Equal Access of Asian Americans to Highly Selective Universities*, 98 YALE L.J. 659, 670-74 (1989); *see also* Kidder, *supra* note 15, at 59-67; John D. Lamb, *The Real Affirmative Action Babies: Legacy Preferences at Harvard and Yale*, 26 COLUM. J.L. & SOC. PROBS. 491, 502-06 (1993).

137. *See* WILLINGHAM & COLE, *supra* note 132, at 84 tbl.3.2; Richard J. Coley, *Differences in the Gender Gap: Comparisons Across Racial/Ethnic Groups in Education and Work*, ETS POLICY INFORMATION REPORT 18-25 (2001); Linda F. Wightman, *Analysis of LSAT Performance and Patterns of Application for Male and Female Law School Applicants*, LSAC RESEARCH REPORT NO. 94-02, at 25 tbl.8 (1994).

138. *See* WILLINGHAM & COLE, *supra* note 132, at 128-38; Dana Keller et al., *Relationships Among Gender Differences in Freshman Course Grades and Course Characteristics*, 85 J. EDUC. PSYCHOL. 702 (1993).

139. *See, e.g.*, William C. Kidder, *Portia Denied: Unmasking Gender Bias on the LSAT and Its Relationship to Racial Diversity in Legal Education*, 12 YALE J.L. & FEMINISM 1 (2000); David K. Leonard & Jiming Jiang, *Gender Bias and the College Predictions of the SAT: A Cry of Despair*, 40 RESEARCH IN HIGHER EDUC. 375 (1999); Susan Sturm & Lani Guinier, *The Future of Affirmative Action: Reclaiming the Innovative Ideal*, 84 CAL. L. REV. 953, 992-97 (1996); Espinoza, *supra* note 54, at 127-38; Andrea L. Silverstein, Note, *Standardized Tests: The Continuation of Gender Bias in Higher Education*, 29 HOFSTRA L. REV. 669 (2000); Katherine Connor & Ellen J. Vargyas, *The Legal Implications of Gender Bias in Standardized Testing*, 7 BERKELEY WOMEN'S L.J. 13 (1992); Kary L. Moss, *Standardized Tests as a Tool of Exclusion: Improper Use of the SAT in New York*, 4 BERKELEY WOMEN'S L.J. 230 (1989); PHYLLIS ROSSER, THE SAT GENDER GAP: IDENTIFYING THE CAUSES (1989); James W. Loewen et al., Gender Bias in SAT Items (1988) (paper presented at the AERA Conference, available at U.S. Dept. of Education, ERIC document # ED 294 915).

possible to simultaneously moderate racial/ethnic item bias and gender item bias.[140] This "win-win" scenario is, in part, a statistical byproduct of the fact that 59% of African American and 58% of Latino SAT test-takers are women (compared to 54% of Whites)—meaning that efforts to moderate gender impact will necessarily reduce racial/ethnic impact to some degree, and vice versa.[141]

2. *How Much Can the* Golden Rule *Approach Reduce the Test Score Gap?*

In analyzing the GRE, Hackett et al. were able to decrease the Black-White test score gap by 18%-33% using item moderation techniques.[142] In a 1998 study, Martha Stocking and other ETS researchers studied impact moderation on the SAT Math section on over 600 items administered to 2.5 million test-takers.[143] Stocking et al. were able to reduce about 20% of the gender gap while also decreasing the Black-White gap by 9%.[144] According to Stocking et al.'s 2001 study of the SAT, the "test construction" method (which yielded more consistent results) reduced 3%-19% of the Black-White Verbal gap, 6%-11% of the Black-White Math gap, 7%-25% of the Latino-White Verbal gap, and 0%-12% of the Latino-White Math gap, at the same time that gender gaps were also lessened.[145]

However, there is reason to view ETS "in house" experimental efforts at impact moderation with some skepticism.[146] For example, in Stocking et al.'s studies, the Verbal items were subject to sixty-four constraints on test content and statistical properties in addition to consideration of impact, and Math items were subjected to 196 such constraints.[147] Such statistical straitjacketing will lessen the

---

140. *See* Stocking et al., *supra* note 124, at 15 tbl.4.; Martha L. Stocking et al., *An Investigation of the Simultaneous Moderation of Average Gender and African-American Score Differences on a Test of Mathematical Reasoning*, ETS RESEARCH REPORT NO. 98-46, at 36 (1998).

141. *See* Stocking et al., *supra* note 124, at 30; *see also* Coley, *supra* note 138, at 20 (illustrating through graphs that American Indians, African Americans, Chicanos, Puerto Ricans, and other Latinos all have higher proportions of female SAT test-takers than Whites).

142. *See* Hackett et al., *supra* note 84, at 27.

143. *See generally* Martha L. Stocking et al., *supra* note 141.

144. *See id.* at 36.

145. *See* Stocking et al., *supra* note 124, at 15 tbl.4.

146. *See* Shapiro et al., *supra* note 79, at 254 (noting the low priority given to impact reduction by ETS regarding the post-*Golden Rule* Illinois Insurance Exam).

147. *See* Stocking et al., *supra* note 124, at 11.

effectiveness of impact moderation, causing ETS estimates to be on the low side. Outside researchers have estimated, for example, that *Golden Rule*-style techniques could decrease the Black-White disparity on the SAT by about 33%-40%.[148] Based on the findings by ETS researchers, as well as outside scholars, we conclude that reducing approximately one-quarter of the Black-White and Chicano-White SAT score gaps is a reasonable goal using item impact reduction techniques.

The meaning of a one-quarter reduction should not be underestimated. To place things in perspective, it is helpful to examine how much or little racial/ethnic disparities have decreased in the last two decades on several standardized tests. Since 1980, African American and Latino high school seniors made gains of about 0.2 standard deviations (relative to Whites) on the National Assessment of Educational Progress (NAEP) Math test, Blacks and Chicanos improved about 0.2 standard deviations on the ACT Math test, and on the SAT Math section Chicano scores remained unchanged and African American performance improved only 0.1 standard deviations.[149] Therefore, the magnitude of impact reduction using *Golden Rule*-style techniques could easily exceed the meager SAT gains made by students of color on the SAT over the past twenty years.

IV. LEGAL ANALYSIS

This section analyzes the law governing standardized tests and Title VI disparate impact claims, including the prospects of enforcing the U.S. Department of Education disparate impact regulations through section 1983 of the Civil Rights Act of 1871. This section also examines the possibility of lodging complaints with the Office for Civil Rights. Because ETS and similar test producers are not recipients of federal financial assistance and are not subject to these civil right statutes, suing colleges and universities on a disparate impact theory

148. Loewen, *supra* note 83, at 86.

149. *See* George Madaus & Marguerite Clarke, *The Adverse Impact of High-Stakes Testing on Minority Students: Evidence from One Hundred Years of Test Data*, *in* RAISING STANDARDS OR RAISING BARRIERS? INEQUALITY AND HIGH-STAKES TESTING IN PUBLIC EDUCATION 85, 89-92 (Gary Orfield & Mindy L. Kornhaber eds., 2001); *see also* David W. Grissmer, *The Continuing Use and Misuse of SAT Scores*, 6 PSYCHOL., PUB. POL'Y, & L. 223, 225 (2000). We included examples from several standardized tests because the pool of students taking the SAT is not representative of all high school students and the self-selectivity of this pool changes over time; these two facts make it difficult to draw firm conclusions about SAT group performance differences over time.

over their use of the SAT in admissions is the only judicial remedy.[150]

A. *Discriminatory Intent: A Dead-end for Plaintiffs*

In the absence of a history of de jure segregation at a particular educational institution, it is difficult for plaintiffs to prevail on an Equal Protection claim against a university for relying on the SAT. To demonstrate an Equal Protection violation on the basis of racial discrimination requires a showing that the state actor was motivated by a discriminatory purpose or intent.[151] Racial discrimination in standardized testing based upon facially-neutral test development procedures does not rise to the level of discriminatory purpose. For example, in *Personnel Administrator of Massachusetts. v. Feeney*,[152] the Supreme Court stated: "'Discriminatory purpose,' however, implies more than intent as volition or intent as awareness of consequences. It implies that the decisionmaker . . . selected or reaffirmed a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group."[153]

In *Village of Arlington Heights v. Metropolitan Housing Development*,[154] the Court specified a non-exhaustive list of factors that can support a finding of discriminatory purpose: (1) the historical background of the policy, particularly if it reveals a series of official actions taken for invidious purposes; (2) the specific sequence of events leading up to the challenged policy; (3) departures from normal

150. *See* Nat'l Collegiate Athletic Ass'n. v. Smith, 525 U.S. 459 (1999) (ruling that the mere fact that the NCAA received funds from schools that, in turn, received federal financial assistance, does not expose the NCAA to lawsuits pursuant to Title IX of the Education Amendments of 1972); Cureton v. Nat'l Collegiate Athletic Ass'n, 198 F.3d 107, 114-19 (3d Cir. 1999) (holding in part that Title VI disparate impact regulations are program specific, and finding that the NCAA could not be sued over the alleged disparate impact of its minimum SAT score eligibility requirement under the theory that the NCAA has "controlling authority" over colleges and is therefore an indirect recipient of federal assistance). *Cf.* Silverstein, *supra* note 134, at 690-91 ("To date there have been no cases which challenge the mere existence of the SAT, and no suits against a university for using the SAT as a decisive factor in its admissions decisions. . . . ETS may not be considered an educational program because it does not specifically receive federal financial assistance.").

151. *See* Washington v. Davis, 426 U.S. 229, 242 (1976) (proof of Equal Protection Clause violations require evidence of discriminatory intent); Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 270 (1977) (ruling that proof of discriminatory purpose or intent is a prerequisite for establishing a constitutional violation).

152. 442 U.S. 256 (1979).

153. *Id.* at 258.

154. 429 U.S. 252 (1977).

procedural sequences; (4) substantive departures, particularly if the factors usually considered important by the decision maker strongly favor a policy contrary to the one implementation; and (5) the legislative or administrative history, especially where there are contemporary statements by members of the decision making body.[155]

As a practical matter, discriminatory purpose is an exceedingly difficult burden of proof in the higher education/standardized testing context. For example, in *United States v. Fordice*,[156] the Supreme Court held that Mississippi's use of ACT cut-off scores in admissions was constitutionally suspect because it was originally adopted just days after the court ordered the University to admit African American students.[157] Further, the standardized test requirement that was traceable to that decision continued to have segregative effect, and Mississippi failed to demonstrate that sole reliance on test scores was educationally necessary.[158]

B. *Title VI Disparate Impact Regulations*

Another litigation option is Title VI of the Civil Rights Act of 1964. The Supreme Court ruled, in *Regents of the University of California v. Bakke*, that Title VI prohibited only the same forms of purposeful discrimination that are forbidden by the Equal Protection Clause.[159] Similarly, in *Guardians Association v. Civil Service Commission of the City of New York*, five justices (in four separate opinions) held that *Bakke* compelled, as a matter of stare decisis, that the terms of section 601 of Title VI required proof of discriminatory intent.[160] The *Bakke/Guardians* Title VI intentional discrimination requirement has not been overturned in subsequent cases.[161]

---

155. *See Vill. of Arlington Heights*, 429 U.S. at 267-68 (1977).

156. 505 U.S. 717 (1992).

157. *See id.*; WILLIAM C. KIDDER, TESTING THE MERITOCRACY: STANDARDIZED TESTING AND THE RESEGREGATION OF LEGAL EDUCATION chap. 2 (book manuscript under submission with Stanford University Press) (noting that Mississippi colleges adopted the ACT requirement one week after Meredith v. Fair, 298 F.2d 696 (5th Cir. 1962)).

158. *See Fordice*, 505 U.S. at 735-39; *see also* Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518, 1530-31 (M.D. Ala. 1991) (rejecting the use of the ACT as the sole criteria for admission to a teacher training program).

159. *See* 438 U.S. 265, 287 (1978).

160. *See* Guardians Ass'n v. Civil Serv. Comm'n, 463 U.S. 582, 610-11 (1983) (Powell, J., concurring); *id.* at 612, (O'Connor, J., concurring); *id.* at 641-42 (Stevens, Brennan, & Blackmun, JJ., dissenting).

161. *See* Alexander v. Choate, 469 U.S. 287, 293 (1985) (stating, in dicta, "Title VI

While the efficacy of Title VI itself is limited by the same discriminatory purpose requirement as the Equal Protection Clause, the U.S. Department of Education regulations *interpreting* Title VI[162] prohibit both intentional discrimination and criteria or practices that have an unwarranted disparate impact on a protected class.[163] Equally important, a different majority in *Guardians* stated that notwithstanding the fact that Title VI requires proof of intentional discrimination, a party can bring a colorable disparate impact claim (at least for limited injunctive and declaratory relief) under Title VI regulations.[164] Thus, federal courts allow plaintiffs to enforce the Department of Education regulations by bringing Title VI claims alleging disparate impact discrimination.[165] The Supreme Court, in

---

itself directly reach[es] only instances of intentional discrimination."); United States v. Fordice, 505 U.S. 717, 732 n.7 (1992) (holding, in a suit brought under both the Equal Protection Clause and Title VI: "Our cases make clear, and the parties do not disagree, that the reach of Title VI's protection extends no further than the Fourteenth Amendment . . . . We thus treat the issues in these cases as they are implicated under the Constitution.").

162. 34 C.F.R. § 100.3(vii)(2) (Lexis 2002).

163. 34 C.F.R. § 100.3(vii)(2) provides:

> A recipient, in determining the types of services, financial aid, or other benefits, or facilities which will be provided under any such program, or the class of individuals to whom, or the situations in which, such services, financial aid, other benefits, or facilities will be provided under any such program, or the class of individuals to be afforded an opportunity to participate in any such program, *may not*, directly or through contractual or other arrangements, *utilize criteria or methods of administration which have the effect of subjecting individuals to discrimination* because of their race, color, or national origin, or have the effect of defeating or substantially impairing accomplishment of the objectives of the program as respect individuals of a particular race, color, or national origin.

34 C.F.R. § 100.3(vii)(2) (emphasis added); *see also* Linda Hamilton Krieger, *Civil Rights Perestroika: Intergroup Relations After Affirmative Action*, 86 CAL. L. REV. 1251, 1299-1300 (1998) (discussing disparate impact and Department of Education's Title VI regulations).

164. *See Guardians*, 463 U.S. at 607 n.27.

165. *See* Krieger, *supra* note 164, at 1300 (citing Villanueva v. Carere, 85 F.3d 481, 486 (10th Cir. 1996); New York Urban League, Inc. v. New York, 71 F.3d 1031, 1036 (2d Cir. 1995); Chicago v. Lindley, 66 F.3d 819, 827 (7th Cir. 1995); Elston v. Talledega County Bd. of Educ., 997 F.2d 1394, 1406 (11th Cir. 1993); David K. v. Lane, 839 F.2d 1265, 1274 (7th Cir. 1988); Gomez v. Illinois State Bd. of Educ., 811 F.2d 1030, 1044 (7th Cir. 1987); Latinos Unidos de Chelsea En Accion (LUCHA) v. Sec'y of Hous. and Urban Dev., 799 F.2d 774, 795 (1st Cir. 1986); United States v. LULAC, 793 F.2d 636, 648 (5th Cir. 1986); Larry P. v. Riles, 793 F.2d 969, 981, (9th Cir. 1986), *as amended on denial of reh'g and reh'g en banc*; Castaneda v. Pickard, 781 F.2d 456, 466 (5th Cir. 1986); Georgia State Conference of Branches of NAACP v. Georgia, 775 F.2d 1403, 1416 (11th Cir. 1985); Young v. Montgomery County Bd. of Educ., 922 F. Supp. 544 (M.D. Ala. 1996); Ass'n of Mexican-American Educators v. California, 836 F. Supp. 1534, 1545 (N.D. Cal. 1993); Grimes v. Sobol, 832 F. Supp. 704, 709 (S.D.N.Y. 1993); Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518, 1522 (M.D. Ala 1991); Theresa P. v. Berkeley Unified Sch. Dist., 724 F. Supp. 698, 716 (N.D. Cal. 1989)).

Alexander v. Choate, noted in dicta that agency regulations designed to implement Title VI can be premised upon a disparate impact theory.[166]

However, Title VI disparate impact regulations were recently dealt a severe blow. In *Alexander v. Sandoval*, the Supreme Court ruled that there is no private right of action to bring a disparate impact suit to enforce Title VI regulations.[167] This was a marked departure from what had been a settled body of jurisprudence, including the unanimous view of the nine circuit courts that addressed the issue.[168] In *Sandoval*, a class action suit challenging Alabama's English-only written driver's license examination policy, the majority found that section 601[169] of Title VI does not authorize a private right of action in disparate impact suits because, under *Bakke* and *Guardians*, section 601 only proscribes intentional discrimination.[170] Next, the *Sandoval* Court found that the legislative intent behind section 602[171] of Title VI was merely to authorize federal agencies to effectuate rights already created under section 601,[172] from which the Court concluded that there was no evidence of congressional intent to create a private right of action to enforce Title VI disparate impact regulations.[173]

C. *Enforcing Disparate Impact Regulations via Section 1983*

Nonetheless, the *Sandoval* Court's ruling did not necessarily sound the death knell for all privately filed Title VI-inspired disparate impact claims. As Justice Stevens noted in dissent:

---

166. *See* 469 U.S. 287, 293-95 (1985) (discussing Guardians Ass'n. v. Civil Serv. Comm'n, 463 U.S. 582 (1983)).

167. 532 U.S. 275 (2001); *see Leading Cases*, 115 HARV. L. REV. 497 (2001) (discussing Alexander v. Sandoval); *see also* Adele P. Kimmel et al., *The Sandoval Decision and Its Implications for Future Civil Rights Enforcement*, FLA. BAR J., Jan. 2002, at 24.

168. *See Sandoval*, 532 U.S. at 295 n.1 (Stevens, Souter, Ginsburg, & Breyer, JJ., dissenting). (summarizing prior cases that expressly or impliedly allowed a private right of action for claims based upon disparate impact).

169. *See* 42 U.S.C. § 2000d (2002) (providing that no person shall, "on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity" covered by Title VI).

170. *See Sandoval*, 532 U.S. at 280-85.

171. 42 U.S.C. § 2000d-1 (2002) (authorizing federal agencies "to effectuate the provisions of [section 601] . . . by issuing rules, regulations, or orders of general applicability.").

172. *See Sandoval*, 532 U.S. at 288-89.

173. *See id.* at 291.

> [T]o the extent that the majority denies relief to the respondents merely because they neglected to mention 42 U.S.C. § 1983 in framing their Title VI claim, this case is something of a sport. Litigants who in the future wish to enforce the Title VI regulations against state actors in all likelihood must only reference § 1983 to obtain relief.[174]

In fact, two viable options will be assessed in this section: bringing section 1983 actions to enforce Department of Education regulations and filing administrative complaints directly with the Department of Education. A third option—congressional repudiation of *Sandoval* akin to the way that the 1991 Civil Rights Restoration Act[175] reined in the Court's decision in *Wards Cove Packing Co., Inc. v. Antonio*[176]—may be equally or more promising. In the short term, however, Republicans control the House of Representatives and the executive branch, and will control the Senate in the upcoming term, which would make passage of such a bill unlikely. Since our expertise is not in politics, we leave it for others to assess legislative solutions in greater depth.

Section 1983 originated with the Civil Rights Act of 1871, a statute intended to enforce Fourteenth Amendment protections amidst efforts by the Ku Klux Klan and other southern White supremacists to deprive Blacks of their nascent rights after the Civil War.[177] Section 1983 states:

> Every person who, under color of any statute, ordinance, regulation, custom, or usage, or any State or Territory of the District of Columbia, subjects, or causes to be subjected, any citizen of the United States or other person within the jurisdiction thereof to the deprivation of any rights, privileges, or immunities secured by the Constitution and laws, shall be liable to the party injured in an action at law, suit in equity, or other proper proceeding for redress.[178]

The crucial phrase "and laws" was added by the Committee on

---

174. *Id.* at 299-300.

175. Civil Rights Act of 1991, Pub. L. No. 102-166, § 1745, 105 Stat. 1071 (1991).

176. 490 U.S. 642 (1989). For further discussion of *Wards Cove*, *see infra* Part IV.D.2.

177. *See* Todd E. Pettys, *The Intended Relationship Between Administrative Regulations and Section 1983's "Laws,"* 67 GEO. WASH. L. REV. 51, 55-56 (1998); *see* Peggy Davis, *Neglected Voices*, *at* http://www.law.nyu.edu/davisp/neglectedvoices/KlanActSpeeches.html (last visited June 12, 2002) (posting the speeches of African American members of the Reconstruction Congress who supported the Civil Rights Act of 1871).

178. 42 U.S.C. § 1983 (1994).

Revision of the Laws, an ambitious effort to consolidate federal statutes that was ratified in 1874.[179] Section 1983 lay dormant as a civil rights tool until the 1960s, when the Supreme Court held, in *Monroe v. Pape*, that section 1983 provides federal remedies against state officials who violate federal rights.[180] Two decades later, in the pivotal case of *Maine v. Thiboutot*, the Court applied a plain meaning test to the phrase "and laws," ruling that section 1983's reach extends to violations of rights protected under any federal law, not just equal protection laws.[181]

Shortly after the *Thiboutot* decision, the Court laid down two limiting principles for courts to apply to section 1983 claims: (1) a plaintiff must establish that he or she is asserting an enforceable "right" which is encompassed by section 1983; and (2) that Congress did not intend to preempt enforcement of section 1983 remedies for a statute by virtue of other comprehensive enforcement mechanisms.[182]

As to the issue of litigating a university's unwarranted reliance on the SAT, the key question is whether the Department of Education's Title VI disparate impact regulations[183] can be privately enforced via section 1983. This issue has yet to be squarely addressed by the Supreme Court, but the prospects of using section 1983 to enforce Title VI disparate impact regulations are dimming. In *Sandoval*, the Court assumed for purposes of deciding the Title VI private right of action issue that regulations promulgated pursuant to section 602 may prohibit disparate impact discrimination.[184] Yet, the *Sandoval* majority questioned in dicta whether it is sound to allow Title VI agency regulations to prohibit disparate impact when such conduct is not itself

179. *See* Cass Sunstein, *Section 1983 and the Private Enforcement of Federal Law*, 49 U. CHI. L. REV. 394, 401-09 (1982); Pettys, *supra* note 178, at 57-60; Lisa E. Key, *Private Enforcement of Federal Funding Conditions Under S 1983: The Supreme Court's Failure to Adhere to the Doctrine of Separation of Powers*, 29 U.C. DAVIS L. REV. 283, 302-06 (1996).

180. *See* 365 U.S. 167, 173-74 (1961).

181. *See* 448 U.S. 1 (1980); *see* Key, *supra* note 180, at 308-13 (giving a defense of the plain meaning test as applied to § 1983).

182. *See* Wilder v. Virginia Hosp. Ass'n, 496 U.S. 498 (1990); *see also* Pennhurst State Sch. and Hosp. v. Halderman, 451 U.S. 1 (1981); Middlesex County Sewerage Auth. v. Nat'l Sea Clammers Ass'n. 453 U.S. 1, 19 (1981); Wright v. City of Roanoke Redevelopment and Hous. Auth., 479 U.S. 418, 423-24 (1987).

183. *E.g.*, 34 C.F.R. § 100.3(vii)(2) (Lexis 2002).

184. *See* Alexander v. Sandoval, 532 U.S. 275, 281, 286 (2001); *see also* Charles F. Abernathy, *Title VI and the Constitution: A Regulatory Model for Defining "Discrimination,"* 70 GEO. L.J. 1 (1981) (arguing that Congress clearly established rights against disparate impact discrimination in section 602 by virtue of its delegation of the definition of discrimination to the administrative agencies responsible for implementing affected programs).

outlawed by Title VI.[185] More writing on the wall appeared in *Gonzaga University v. Doe*, in which the Court held that the Family Educational Rights and Privacy Act (FERPA) could not be privately enforced through section 1983, and declared, "We now reject the notion that our cases permit anything short of an unambiguously conferred right to support a cause of action brought under § 1983."[186] However, in *Gonzaga University* the Court distinguished FERPA from Title VI and Title IX, which create individual rights because the plain language of these statutes unmistakably focuses on the benefited classes.[187]

In the absence of controlling Supreme Court precedent, it is instructive to compare and contrast the Third and Sixth Circuit approaches to the issue of section 1983 and disparate impact regulations. Until recently, civil rights groups could point to the Third Circuit's decision in *Powell v. Ridge*, in which the court held that there is a private right of action to enforce Title VI disparate impact regulations, and that section 1983 can be used to enforce these regulations.[188] While *Sandoval* unquestionably overruled *Powell* by limiting a private right of action in Title VI suits to intentional discrimination,[189] *Powell*'s section 1983 holding was not disapproved by the Court. The *Sandoval* majority responded with silence to Justice Stevens' comment in the dissent that the availability of section 1983 remedies rendered *Sandoval* "something of a sport."[190]

In *Powell*, the plaintiffs (a coalition of parents and educational organizations) brought a Title VI and section 1983 action against Pennsylvania state officials for declaratory and injunctive relief, alleging that the state's school funding practices had a racially

185. *See* 532 U.S. at 286 n.6 (citing Guardians Ass'n. v. Civil Serv. Comm'n, 463 U.S. 582, 613 (1983)) ("We cannot help observing, however, how strange it is to say that disparate-impact regulations are 'inspired by, at the service of, and inseparably intertwined with' § 601 . . . when § 601 permits the very behavior that the regulations forbid.") (O'Connor, J., concurring); *id.* ("If, as five members of the Court concluded in *Bakke*, the purpose of Title VI is to proscribe *only* purposeful discrimination . . . regulations that would proscribe conduct by the recipient having only a discriminatory effect . . . do not simply 'further' the purpose of Title VI; they go well beyond that purpose.").

186. 122 S.Ct. 2268, 2275 (2002).

187. *See id.* at 2275-76.

188. *See* 189 F.3d 387 (3d Cir. 1999); *see* Bradford C. Mank, *Using § 1983 to Enforce Title VI's Section 602 Regulations*, 49 KAN. L. REV. 321, 365-67 (2001) (commenting on the importance of the Powell v. Ridge section 1983 ruling).

189. *See* Alexander v. Sandoval, 532 U.S. 275 (2001).

190. *Id.* at 299-300.

disparate impact.[191] The Third Circuit, while not reaching the merits of plaintiffs' claims, reversed the lower court's dismissal of the complaint.[192] The court rejected defendant's contention that Title VI regulations were sufficiently comprehensive to preclude section 1983 remedies.[193] Rather, the court was satisfied that the Department of Education's Title VI regulations created a federal right.[194] Moreover, the *Powell* court ruled:

> Neither Title VI nor the Department of Education regulation establishes "an elaborate procedural mechanism to protect the rights of [individual plaintiffs]". . . Nor is it possible to describe the administrative remedies Title VI and the regulations establish as "unusually elaborate". . . Indeed, the statutory scheme under Title VI does not specifically provide individual plaintiffs with any administrative remedy."[195]

In summary, the *Powell* Third Circuit panel found that section 1983 suits are not incompatible with Title VI enforcement regulations.[196]

Yet the promise of *Powell* ebbed quickly. In December 2001, a different Third Circuit panel held, in *South Camden Citizens in Action v. New Jersey Dept. of Environmental Protection*, that because Title VI only prohibits intentional discrimination, plaintiffs do not have a right to enforce EPA Title VI disparate impact regulations via section 1983.[197] The *South Camden* court essentially "Sandovalized" the inquiry into section 1983 as an enforcement mechanism for Title VI disparate impact regulations, ruling that because section 601 of Title VI proscribes only intentional discrimination, section 602 could not authorize agencies to promulgate disparate impact regulations pursuant to Title VI.[198] The *South Camden* majority strained to distinguish *Powell* in order to overrule it without candidly acknowledging that it was ignoring *Powell*'s stare decisis value. The *South Camden* panel declared that *Powell* "should not be overread" and that *Powell* assumed rather than analyzed "the foundation issue that is central here, *i.e.*, whether a regulation in itself can create a right enforceable under

191. *Powell*, 189 F.3d. at 391-92.
192. *See id.* at 405.
193. *See id.* at 401-03.
194. *See id.* at 401.
195. *Id.* at 402 (citing Smith v. Robinson, 468 U.S. 992, 1010-11 (1984); Middlesex County Sewerage Auth. v. Nat'l Sea Clammers Ass'n, 453 U.S. 1, 13 (1981)).
196. *See id.* at 403.
197. *See* 274 F.3d 771 (3d Cir. 2001).
198. *See id.* at 786-90.

section 1983."[199]

A dissenting judge in *South Camden* decried the majority's "analytical alchemy" for confusing the tests for an implied private right of action and for section 1983, as well as for disregarding the binding authority of *Powell* even after the majority acknowledged that the *Powell* court "held" that "a disparate impact discrimination claim could be maintained under section 1983 for a violation of a regulation promulgated pursuant to section 602."[200] The *South Camden* court was incorrect to "Sandovalize" its analysis of section 1983 and Title VI disparate impact regulations; for the reasons we state below, the *South Camden* court applied the wrong standard when it required proof of specific congressional intent to authorize a private right of action to enforce disparate impact regulations via section 1983.[201] Unlike an implied private right of action, section 1983 expressly authorizes a private right of action.[202] Accordingly, the Court noted in *Wilder v. Virginia Hosp. Ass'n* that the question of whether section 1983 can serve as the basis for a suit involves a "different inquiry" than that underlying the question of whether the same statute allows a private right of action.[203]

The *Wilder* Court made this analytical distinction because section 1983 "provides an alternative source of express congressional authorization of private suits . . . these separation-of-powers concerns are not present in a section 1983 Case."[204] In contrast to section 1983, whether or not there is an implied private right of action is a question that implicates separation of powers in two respects. First, Article III of the Constitution proscribes that lower federal courts may only review those matters that Congress has statutorily granted jurisdiction, meaning that courts risk encroaching upon a congressional function when they allow an implied private right of action to form the basis for jurisdiction.[205] In addition to this danger of judicial lawmaking, private

199. *Id.* at 784.

200. *Id.* at 791-95 (McKee, J., dissenting).

201. *See* Mank, *supra* note 189 at 353-59; Brief of Amici Curiae Law Professors Concerned About Environmental Justice, South Camden Citizens in Action v. New Jersey Dept. of Envtl. Prot., 274 F.3d 771 (3d Cir. 2001) (No. 01-224 & 01-2296).

202. *See supra* note 179 and accompanying text (quoting 42 U.S.C. section 1983 (1994)).

203. *See* 496 U.S. 498, 508 n.9 (1990).

204. *Id.*

205. *See* Key, *supra* note 180, at 299; Mank, *supra* note 189, at 354; *see* Sunstein, *supra* note 180, at 415.

rights of action also invoke separation of powers concerns because Congress alone has the power to interfere with states' lawmaking powers, as it is the only branch of the federal government in which states are represented.[206]

In light of the absence of such serious separation of powers implications, the Supreme Court, in *Blessing v. Freestone*, *Wilder,* and other cases, applied a less stringent three-part test to assess when a statute creates an enforceable right actionable under section 1983.[207] In *Blessing*, a unanimous Court reiterated the three traditional factors: (1) the plaintiff must be an intended beneficiary of the statute; (2) the plaintiff's interests must not be so "vague and amorphous" that they extend beyond the judiciary's sphere of competence; and (3) a statute must clearly impose a binding obligation on the States, as evidenced by mandatory, not precatory terms.[208] Satisfaction of this test creates the rebutable presumption that there is a right enforceable under section 1983.[209] The presumption of a right can be rebutted by either express language in the statute itself precluding section 1983, or by evidence that Congress impliedly forbid section 1983 because it created a comprehensive enforcement scheme that is incompatible with section 1983 individual remedies.[210] *Gonzaga University v. Doe*[211] did not change the three-part *Blessing* test, nor did it expressly overrule *Wilder*.

The *Blessing* test led to *Loschiavo v. City of Dearborn*, where the Sixth Circuit ruled that section 1983 can be a mechanism for enforcing rights created by federal regulations.[212] In *Loschiavo*, the court held that Federal Communications Commission (FCC) regulations preempted local zoning ordinances, finding that the three-part test was satisfied and that since federal regulations carry the force of law, regulations may create enforceable rights.[213] The *Loschiavo* precedent

206. *See* Key, *supra* note 180, at 299; Richard W. Creswell, *The Separation of Powers Implications of Implied Rights of Action,* 34 MERCER L. REV. 973, 991-92 (1983).

207. *See* Blessing v. Freestone, 520 U.S. 329, 340-41 (1997); Wilder v. Virginia Hosp. Ass'n, 496 U.S. 498, 509 (1990); Livada v. Bradshaw, 512 U.S. 107, 132-34 (1994); *see also* Golden State Transit Corp. v. City of Los Angeles, 493 U.S. 103, 107-08 (1989).

208. *See Blessing*, 520 U.S. at 340-41.

209. *See id.*

210. *See id.* at 341.

211. 122 S.Ct. 2268 (2002); *Cf. id.* at 2285-86 (Stevens & Ginsburg, JJ., dissenting) (arguing that, despite assurances to the contrary, the majority eroded the principle that rights under section 1983 are presumptively enforceable).

212. 33 F.3d 548 (6th Cir. 1994).

213. *See id.* at 551-53.

has led to two recent district court rulings within the Sixth Circuit allowing plaintiffs to bring section 1983 actions to enforce rights contained in Title VI disparate impact regulations.

The post-*Sandoval* case of *White v. Engler* is particularly relevant to our analysis of the SAT, as it involved a disparate impact challenge to the practice of awarding merit scholarships based upon the Michigan Education Assessment Program High School Test (MEAP Test).[214] In *Engler*, although the district court did not reach the merits of plaintiffs' challenge to the MEAP test, the court denied defendants' motion to dismiss because the Department of Education's disparate impact regulations unambiguously imposed a binding obligation on the states. The court found that the regulations were clearly intended to benefit the African Americans, Hispanics, and Native Americans who brought suit, and that the regulations were unquestionably within the province of judicial competence.[215]

The other relevant post-*Sandoval* district court case in the Sixth Circuit is *Lucero v. Detroit Public Schools*, in which plaintiffs moved for a preliminary injunction to prevent a new Detroit elementary school (with an overwhelmingly African American and Latino student population) from opening on a site allegedly contaminated by industrial waste.[216] While denying plaintiffs' motion on other grounds, the district court ruled that plaintiffs satisfied all three prongs of the *Blessing/Wilder* test and could enforce Title VI disparate impact regulations via section 1983.[217]

D. *The SAT: Proving the Elements of a Disparate Impact Claim*

After overcoming the private enforcement hurdle through section 1983, the actual requirements for establishing a disparate impact case are relatively straightforward. In Title VI disparate impact analysis, the plaintiffs bear the initial burden of establishing that the challenged test or test use has a demonstrated disparate impact by race and ethnicity. After this prima facie showing has been made, it is defendant's burden of proof to establish that the challenged test or test use is educationally

214. *See* White v. Engler, 188 F. Supp. 2d 730 (E.D. Mich. 2001).

215. *See id.* at 744.

216. *See* 160 F. Supp. 2d 767, 772-73 (E.D. Mich. 2001) (explaining the University of Michigan's environmental study of the site).

217. *See id.* at 781-84.

justified. If the defendant meets this burden, plaintiffs may still prevail upon a disparate impact theory if plaintiffs can convince the court that there is an equally effective and less discriminatory alternative.[218] This three-part burden-shifting framework mirrors the requirements for Title VII employment discrimination disparate impact cases.[219] Courts confronting Title VI disparate impact challenges therefore often rely on Title VII cases, particularly since the case law is much more extensive in the employment context.[220] There is a paucity of Title VI standardized testing cases challenging college and university admission practices.[221] This may be a reflection of the availability of affirmative action as a counterbalance to disparate impact,[222] and it may also reflect a recognition on the part of plaintiffs' attorneys that Title VI disparate impact cases are difficult to win and may have even less viability in the

218. *See* U.S. DEP'T OF EDUC. OFFICE FOR CIVIL RIGHTS, *supra* note 92, at 54-58 (2000) (summarizing Title VI disparate impact analysis).

219. In a piece that came out while this article was at the final edit stage, Jennifer Braceras argues that the Title VII disparate impact framework should not be applied to Title VI standardized testing claims. Braceras, *supra* note 43, at 1177-1203. Rather than proving educational necessity, which she terms a "charade," Braceras urges reforms to eliminate unfair questions or confining the analysis of test bias to the "totality of circumstances" inquiry in an intentional discrimination claim. *See id.* at 1180. For the reasons discussed in Parts II and III, we conclude both that test developers have consistently resisted efforts to reduce item impact through *Golden Rule*-style procedures despite evidence that such techniques are workable, and that conventional methods of flagging biased items (DIF) are irrevocably flawed. We therefore conclude that Title VI disparate impact litigation is an important tool for addressing a serious problem that will not, in all likelihood, be rectified otherwise. Similarly, as indicated by our discussion of United States v. Fordice, 505 U.S. 717 (1992), *infra* Part IV(a), the prospects of bringing successful intentional discrimination claims under Title VI or the Equal Protection Clause over the use of educational standardized tests are exceptionally meager unless the offending institution has a diehard segregationist history. We therefore conclude that Bracera's recommendations have a "let them eat cake" quality; foreclosing the availability of disparate impact analysis would preclude legal remedies precisely where they are most needed.

220. *See infra* Part IV.D.1-2.

221. *See* Krieger, *supra* note 164, at 1300-01. Krieger reports:

> Although various lower federal courts have followed Guardians [sic] and permitted Title VI plaintiffs to proceed under a disparate impact theory in actions to enforce the regulations, no reported case has ever challenged the use of either the SAT, the LSAT, the Graduate Record Exam (GRE), or the Medical College Admissions Test (MCAT). Indeed, as of the writing of this Article, I have been unable to find a single reported Title VI or Title IX case in which college or graduate school admissions criteria have been challenged. Thus, unlike employers, whose selection procedures have for years been subject to challenge under Title VII, institutions of higher education have been left to define and assess merit in admissions decision making in an atmosphere utterly devoid of legal contest.

Id. (internal citation omitted).

222. *See* Miranda Massie, *A Student Voice and a Student Struggle: The Intervention in the University of Michigan Law School Case*, 12 LA RAZA L.J. 231, 233 (2001).

future.[223]

1. *Determining Disparate Impact*

In the Title VII employment context, the Supreme Court declared that there is "no rigid mathematical threshold" to overcome a facially neutral practice as long as statistical disparities are sufficiently large to raise an inference that the challenged practice caused the disparate results.[224] Courts have essentially adopted the same requirement for Title VI disparate impact claims.[225] Plaintiffs' initial prima facie burden of establishing disparate impact is usually less onerous than contesting educational necessity or providing a workable less discriminatory alternative.[226]

One recognized benchmark for assessing disparate impact is the Equal Employment Opportunity Commission's "Four-Fifths Rule," which allows a court to find an adverse impact when the passing rate for the minority group is less than 80% of the passing rate for the majority group (Whites).[227] While results would vary depending on factors such as the particular test use involved, the appropriate applicant pools, and the level of selectivity, application of the Four-Fifths Rule would, in a majority of cases, allow plaintiffs to establish their initial disparate impact burden in a post-affirmative action environment where the SAT was an influential admissions factor. For example, the Black-White gap on the SAT is generally about one standard deviation.[228] In the extreme example of a university that used the SAT as the sole criteria for admission, and with a one standard deviation gap (assuming a normal distribution and that applicants fairly represented the larger population), if 25% of Whites were admitted,

223. *See* Krieger, *supra* note 166, at 1301 ("The dearth of activity under Title VI may, among other things, reflect a lack of confidence in the viability of the Guardians rule.").

224. *See* Watson v. Fort Worth Bank and Trust, 487 U.S. 977, 994-95 (1988); *see* Wards Cove Packing Co., Inc. v. Antonio, 490 U.S. 642, 656-57 (1989).

225. *See, e.g.*, Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518, 1523-29 (M.D. Ala. 1991) (adopting Title VI disparate impact requirements in a challenge to the use of the ACT for a teacher training program); GI Forum v. Texas Educ. Agency, 87 F. Supp. 2d 667, 677-78 (W.D. Tex. 2000) (adopting Title VI disparate impact requirements in challenge to the Texas Assessment of Academic Skills, a standardized test required for high school graduation).

226. *See Watson*, 487 U.S. at 994 (noting that establishing disparate impact is "relatively easy" when appropriate statistical proof is proffered).

227. *See* 29 C.F.R. § 1607 (Lexis 2002). *See also GI Forum*, 87 F. Supp. 2d at 675-76, 678 (accepting the 80% rule as an appropriate measure in a Title VI standardized testing case); Groves, 776 F. Supp. at 1526.

228. *See supra* Part II.

only about 5% of African Americans would be admitted.[229] Under this hypothetical worse-case scenario, plaintiffs could easily meet their prima facie burden since the Black acceptance rate is a mere 20% of the White acceptance rate.[230]

A second recognized test for identifying statistical disparities for adverse impact purposes is the so-called "Shoben formula," or "z-score" statistic, which involves calculating the differences between independent proportions.[231] Whereas the Four-Fifths Rule is an intuitive guidepost, Professor Shoben's z-score statistic is a more reliable method of accounting for differences in sample size and the magnitude of differences in acceptance rates.[232] The z-score technique involves three preconditions (independence, randomness, and sufficiently large sample size)[233] and starts with the null hypothesis that there are no racial and ethnic differences in pass rates in the relevant population.[234] The point of using z-scores or other tests of statistical

---

229. *See* Paul R. Sackett & Steffanie L. Wilk, *Within-Group Norming and Other Forms of Score Adjustment in Preemployment Testing*, 49 AM. PSYCHOL. 929, 942 (1994) (providing this example for the GATB test, which also has a one standard deviation gap).

230. It should be noted that some limited data suggests that the SAT II achievements tests improve admission chances for Latinos and APAs compared to the SAT I because students can take a foreign language test like Spanish, Chinese, or Korean for one of their three SAT II tests. *See* Steven A. Holmes, *SAT II Boosts Diversity, Threatens Controversy*, N.Y. TIMES, July 22, 2001.

231. *See* Groves v. Alabama State Bd. Of Educ., 776 F. Supp. 1518, 1527 (M.D. Ala. 1991); *GI Forum*, 87 F. Supp. 2d at 675-76, 678 (accepting the Shoben formula as an appropriate measure in a Title VI standardized testing case). *See also* Frazier v. Consolidated Rail Corp., 851 F.2d 1447, 1450 n.5 (D.C. Cir. 1988).

232. *See* Elaine W. Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793, 812 (1978). *See also Groves*, 776 F. Supp. at 1527-28 (approvingly citing to Shoben's article and technique).

233. *See* Shoben, *supra* note 233, at 801. Independence is compromised if students can take the test repeatedly or can cheat by passing on test information to subsequent test takers. Randomness is compromised if the self-selected population that applies for a college, takes a test, etc. differs substantially from the larger population. Sample size is adequate if there are at least ten passers and failers in each group when the population is very large. *See id.* at 801.

234. *See id.* at 804. While conservative (and some other) critics might question this assumption as flying in the face of reality, it is important to point out that the assumption is merely an artifact of the Title VI and Title VII burden-shifting framework, and a plaintiff cannot win a case merely by establishing substantial racial/ethnic differences in test scores or admission rates.

For a definition of the null hypothesis, see Thomas J. Campbell, *Regress on Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 STAN. L. REV. 1299 (1984). Professor Campbell explains:

> Null hypotheses are strawmen, established for the purpose of being refuted. In a statistical study, if a researcher suspects that some situation is true, he or she will state the opposite of that situation, run tests under the assumption that this opposite is true, and analyze the results. If the results are that this opposite is

significance is to determine whether there is ample evidence to reject the null hypothesis.[235]

A real example can assist readers in understanding how z-score statistics are utilized to establish disparate impact. UC Berkeley's entering class of 1998 was the first class admitted under California's Proposition 209 and the UC Regents SP-1 Resolution, which banned race-conscious affirmative action in public university admissions.[236] In response, five civil rights organizations soon brought *Rios v. Regents of the University of California*[237] (the lead plaintiff was later changed to *Castañeda*), a class action challenging UC Berkeley admission policies, including allegations that Berkeley placed an unjustified emphasis on SAT scores and unfairly awarded GPA bonus points for honors classes (which affluent White high schools were much more likely to offer than schools with large proportions of African Americans and Latinos).[238] That year, UC Berkeley admitted 28.1% of all applicants (8,438/30,038), including 31.2% of Whites (2,778/8,892), 20.6% of Latinos (647/3139), and 19.3% of African Americans (241/1249).[239]

How would the Shoben formula be applied to the 1998 Berkeley admissions cycle?[240] The first step is to calculate the overall proportion of applicants who were admitted (0.281) and rejected (0.719). These two proportions are then multiplied, and we can label this product "PROD." Here, PROD equals 0.202 (0.281 x 0.719). PROD can then

---

> untrue, and they are so extraordinary that the probability that they are a product of chance is only five percent or less, the researcher will infer that this assumed opposite situation is unlikely.

Id. at 304.

235. *See* David H. Haye & David A. Freedman, *Reference Guide on Statistics* 332, 378-79, *in* FEDERAL JUDICIAL COUNCIL, REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (2002), *available at* http://air.fjc.gov/public/pdf.nsf/lookup/sciman00.pdf/$file/sciman00.pdf.

236. *See supra* note 14.

237. Rios v. Regents of the Univ. of California, Compl. No. 99-0525 (filed in the U.S. District Court, N.D. Cal., Feb 2, 1999) [hereinafter Rios Complaint].

238. *See id.* This action was filed by the Asian Pacific American Legal Center of Southern California, the ACLU, the Lawyers' Committee for Civil Rights of the San Francisco Bay Area, MALDEF, and the NAACP Legal Defense and Education Fund, Inc., on behalf of African American, Chicano/Latino, Native American, and Filipino American applicants. *See also* Lawrence, *supra* note 1, at 942-48; Evelyn Nieves, *Civil Rights Groups Suing Berkeley Over Admission Policy*, N.Y. TIMES, Feb. 3, 1999, at A11; Pamela Burdman, *Lawsuit Against UC Berkeley Claims 'Colorblind' Admissions Policy Is Unjust*, S.F. CHRON., Feb. 3, 1999, at A13.

239. Rios Complaint, *supra* note 238, at 11-12.

240. *See* Shoben, *supra* note 233, at 804.

be used to calculate the Standard Error, which is a "measure of the variability of sample means in a sampling distribution."[241] The Standard Error is equal to the square root of PROD/number of minority group in the sample plus PROD/number of Whites in the sample.[242]

$$\text{Standard Error} = \sqrt{\text{PROD}}\ /\text{N(minority)} + \text{PROD/N(White)}$$

In the above example, the Standard Error equals 0.00933 for Latinos (compared to Whites) and 0.0136 for African Americans (compared to Whites). Lastly, the z-score statistic equals the sample pass rate difference divided by the standard error.[243]

$$Z = \frac{\text{White Pass Rate} - \text{Minority Group Pass Rate}}{\text{Standard Error}}$$

For UC Berkeley's 1998 pool, Z equals 11.36 for Latinos and 8.75 for African Americans. A z-score of 1.96 or higher is needed to reject the null hypothesis with 95% confidence.[244] Thus, in the above example plaintiffs would clearly be able to meet their prima facie disparate impact burden.

Note that in our example, the z-score is higher for Latinos than African Americans even though the Latino-White gap in acceptance rates is smaller than the Black-White gap. This illustrates a crucial distinction between the Shoben formula and the Four-Fifths Rule: with the Shoben statistic, a smaller disparity may still yield a higher z-score if the sample size is much larger, and vice versa (here we had 3,139 Latinos and 1,249 African Americans). The Four-Fifths Rule, which from the beginning was intended as a non-technical guidepost for employers, is less helpful than statistical analysis in that larger sample sizes allow for more precise conclusions about disparate impact, whereas smaller samples require larger disparities to reach statistical significance.[245] Thus, in cases with very small samples, using the Four-Fifths Rule alone can incorrectly suggest a disparate impact; conversely, in cases with large samples, exclusive reliance on the Four-Fifths Rule can obscure the presence of a legally cognizable disparate

241. *Id.* at 802 n.39.
242. *See id.* at 802, 804.
243. *See id.* at 803, 805.
244. *See id.* at 805.
245. *See id.* at 806

impact.[246]

The courts sometimes consider issues of "practical significance" in addition to statistical significance,[247] so both plaintiffs and defendants in Title VI disparate impact cases are better off retaining both technical/psychometric expert witnesses as well as expert witnesses who can place test score disparities in their proper educational, historical, and sociological context. For example, in *Groves v. Alabama State Board of Education*, a challenge to the use of the ACT cut-off for entrance to teacher training programs, the court evaluated both parties' statistical evidence and concluded:

> Here, both the plaintiffs and the State Board have wrapped themselves in complex statistical data and terminology. However, this is one of those rare cases where if one stands back and applies reason and common sense the answer is apparent . . . . [T]he ACT requirement has resulted in substantial adverse racial impact. Indeed, to reach any other conclusion the court would have to close its eyes to the obvious.[248]

2. *Determining Educational Necessity*

After the plaintiff establishes a prima facie case of disparate impact, the defendant has the burden of justifying its use of a standardized test by proffering evidence of "educational necessity." This standard of "educational necessity is similar to "business necessity" in Title VII disparate impact litigation.[249] It is important to

---

246. *See* Shoben, *supra* note 233, at 806-10. *See also* Groves v. Alabama State Bd. Of Educ., 776 F. Supp. 1518, 1527-28 (M.D. Ala. 1991) (reviewing problems courts have encountered when applying the Four-Fifths Rule); Joseph L. Gastwirth, *Employment Discrimination: A Statistician's Look at Analysis of Disparate Impact Claims*, 11 LAW & INEQUALITY 151, 155 (1992) (arguing that statistical testing is preferred to the Four-Fifths Rule).

247. GI Forum v. Texas Educ. Agency, 87 F. Supp. 2d 667, 676 (W.D. Tex. 2000) ("In addition to evaluating the statistical impact of the examination, the Court has, as the behest of both parties, considered the 'practical consequences' or 'practical impact' of the high failure rates of minorities. That consideration involves careful examination of the immediate and long-term effects of the statistically disparate failure rates."); *Groves*, 776 F. Supp. at 1528-29 (discussing practical impact). There is a similar "practical impact" analysis in Title VII. *See* Watson v. Fort Worth Bank and Trust, 487 U.S. 977, 995 n.3 (1988) (O'Connor, J.) (noting that "statistics 'come in infinite variety and their usefulness depends on all of the surrounding facts and circumstances'") (citing Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324, 340 (1977)).

248. *Groves*, 776 F. Supp. at 1529.

249. *See* Board of Educ. v. Harris, 444 U.S. 130, 151 (1979) (holding that defendant's evidence of educational necessity may rebut showing of disparate impact in case involving Emergency School Aid Act); Larry P. v. Riles, 793 F.2d 969, 982 (9th Cir. 1984) (ruling

note that in *Watson v. Fort Worth Bank & Trust*, a case involving alleged racial bias in subjective employment decision-making, four of the eight Supreme Court justices indicated a willingness to substitute "reasonableness" for "business necessity" as the employer's burden.[250] A year later in *Wards Cove Packing Co. v. Antonio*, the majority adopted doctrinal changes suggested in *Watson*, but the ruling did not specify what exactly would be required of the employer in terms of formal validation of their selection procedures.[251] However, *Watson* and *Wards Cove* are no longer controlling regarding Title VII, because the 1991 Civil Rights Act expressly revived the "business necessity" defense in disparate impact cases.[252]

In *Groves v. Alabama State Board of Education*,[253] a disparate impact challenge to the use of ACT cut-off scores for admission to teacher training programs, the district court relied extensively on *Wards Cove*,[254] and it borrowed *Wards Cove*'s lower standard for educational justification.[255] While the *Groves* decision was rendered before the 1991 Civil Rights Act,[256] a more troubling case is the recent district court decision in *GI Forum*, a disparate impact challenge to the Texas Assessment of Academic Skills (TAAS).[257] Despite *Wards*

---

that the defendant carries the burden of establishing the educational necessity of its test use); Debra P. v. Turlington, 644 F.2d 397, 407 (5th Cir. 1981) (same holding).

250. 487 U.S. 977, 997-98 (1988). For further discussion of *Watson*, see Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1229-31 (1995).

251. *See* 490 U.S. 642, 656-61 (1989).

252. *See* Civil Rights Act of 1991, Pub. L. 102-166, § 105, 105 Stat. 1074 (codified at 42 U.S.C. § 2000e-2 (1991)). *See also* David Benjamin Oppenheimer, *Negligent Discrimination,* 141 U. PA. L. REV. 899, 925-30 (1993) (commenting on the 1991 Civil Rights Act and its impact on *Watson*, *Wards Cove*, and related cases); Preston C. Green, *Can Title VI Prevent Law Schools from Adopting Admissions Practices that Discriminate Against African Americans?*, 24 S.U. L. REV. 237, 249-50 (1997) (discussing *Watson*, *Wards Cove*, and the 1991 Civil Rights Act as they pertain to Title VI disparate impact litigation).

253. Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518 (M.D. Ala. 1991).

254. *See id.* at 1523 ("Although both *Georgia State Conference* and *Quarles* articulate the proof necessary to sustain a disparate-impact claim under Title VI's regulations, the Title VII law from which they borrow has since been redefined, particularly by the Supreme Court in *Wards Cove Packing Co. v. Antonio,* 490 U.S. 642, 109 S.Ct. 2115, 104 L.Ed.2d 733 (1989). Accordingly, the court relies principally on *Wards Cove* and other, subsequent Title VII decisions in evaluating plaintiffs' challenge to the ACT requirement in the following sections of this opinion.") (citations omitted).

255. *See Groves*, 776 F. Supp. at 1529-30.

256. *See* Elston v. Talladega County Bd. of Educ., 997 F.2d 1394, 1412 (11th Cir. 1993) (reviewing educational necessity cases and cautioning that *Groves* came out after *Wards Cove* but before the 1991 Civil Rights Act).

257. *See GI Forum*, 87 F. Supp. 667 (W.D. Texas 2000).

Cove's questionable authority, the district court judge in *GI Forum* cited *Wards Cove* as authority for holding that the TAAS test serves the "legitimate educational goals" of the Texas Education Association.[258] While we believe that the district court's use of a less stringent "reasonableness/legitimate goal" standard in *GI Forum* amounted to improper judicial lawmaking in light of the Civil Rights Act of 1991, it remains the case that even a true "educational necessity" standard is not one that is tremendously difficult for defendant universities to satisfy. In reality, the courts have given considerable deference to the professional testing industry and to defendants such as state school boards. It is more accurate to say that the courts require "a substantial legitimate justification" for the practice in question.[259]

In high stakes testing cases under Title VI/Title IX disparate impact, the courts have traditionally looked to the relevant body of evidence regarding the validity, reliability, and fairness of the test and test use.[260] Thus, courts are far more likely to find educational necessity to be lacking where an institution is using a standardized test in a manner inconsistent with the established guidelines of the test producer and the educational measurement establishment.[261] A prime example is the case of Daniel Wurangian, a Latino and Asian high school student who dreamt of attending the U.S. Naval Academy to

258. *See id.* at 679 ("Instead, an educational necessity exists where the challenged practice serves the legitimate educational goals of the institution. *Wards Cove*, 490 U.S. at 659, 109 S.Ct. 2115. In other words, the TEA must merely produce evidence that there is a manifest relationship between the TAAS test and a legitimate educational goal. *Teal*, 457 U.S. at 446. The Court finds that the TEA has met its burden.").

259. *See* Elston v. Talladega County Bd. of Educ., 997 F.2d 1394, 1412 (11th Cir. 1993) ("Under the Title VI disparate impact scheme, once plaintiffs have demonstrated a disparate impact, defendants bear the burden of demonstrating that their challenged practice is supported by a substantial legitimate justification."); Larry P. v. Riles, 793 F.2d 969, 982 n. 9 (9th Cir. 1984) (defining "educational necessity" as proof that a "given requirement has a manifest relationship to the education in question").

260. *See* U.S. DEP'T OF EDUC. OFFICE FOR CIVIL RIGHTS, *supra* note 92, at 57.

261. *See, e.g.*, Groves v. Alabama Bd. Of Educ., 776 F. Supp. 1518, 1531 (M.D. Ala. 1991) (concluding that Alabama State Board of Education's use of a rigid ACT cut-off score for entrance into teacher training programs "fall[s] far outside the bounds of even 'a good faith exercise of professional judgment.'") (citing Richardson v. Lamar County Bd. of Educ., 729 F. Supp. 806, 823 (M.D. Ala. 1989)); Cureton v. Nat'l Collegiate Athletic Ass'n, 37 F. Supp. 2d 687, 707-09, *rev'd on other grounds*, 198 F.3d 107 (3d Cir. 1999) (rejecting the educational necessity of the NCAA's SAT eligibility cut-off score because the SAT has only been validated as a predictor of first-year GPA, not college graduation, and because the NCAA did not demonstrate an independent basis for choosing its cut-off); Sharif v. New York State Educ. Dep't, 709 F. Supp. 345, 362 (S.D.N.Y. 1989) (ruling in a Title IX disparate impact case, that defendants failed to establish a reasonable relationship between the use of the SAT to award scholarships and encouraging high school academic achievement because the "SAT was not designed to measure achievement in high school and was never validated for that purpose").

become a Navy pilot.[262] Wurangian, who served as the battalion commander for his Los Angeles high school in the Naval Junior Reserve graduated with a 3.64 GPA, took the SAT four times and managed to score just over 1000.[263] In a surprisingly frank November 2001 letter from the Naval Academy's head of candidate guidance, Wurangian was informed that he did not score high enough on the SAT to meet the Academy's minimum cut-off, and he was therefore ineligible to receive an application:

> We have carefully evaluated all of the information which you have submitted to date. At the present time, your College Board tests do not indicate sufficient academic achievement for you to be designated an official candidate and receive an application packet. Our pre-qualifying levels for the SAT-I are 530 verbal/570 math and for the ACT are 22 English/24 math. Either test is acceptable for admission. Keep in mind that test scores are the minimum levels needed to receive an application.[264]

The head of the Naval Academy Admissions Office also recommended that Wurangian retake the SAT yet again to raise his scores.[265] The use of such rigid, psychometrically unvalidated cut-off scores runs contrary to the positions of both the College Board and the National Association of College Admission Counseling, of which the Naval Academy is a member.[266] Another legally suspect use of cut-off scores is the state of Florida's requirement that winners of the top level of the Bright Futures scholarship program, which pays 100% of recipients' tuition at public universities, score at least 1270 on the SAT.[267] Whereas about 11% of White students received Bright Futures full scholarships between 1999 and 2001, only 4% of Latinos and 1%

---

262. *See* Diana Jean Schemo, *Spurned Student Challenges Naval Academy on Test Scores*, N.Y. TIMES, July 28, 2002; Ariel Sabar, *Pre-admissions Policy at Academy Challenged: Group Claims School Misuses Test Scores to Keep Some From Applying*, BALTIMORE SUN, July 26, 2002, at 1B; Letter from Christina Perez of FairTest to Vice Admiral Richard Naughton of the Naval Academy (July 23, 2002), *available at* http://www.fairtest.org/pr/Naval_Acc_Letter.html (last visited July 29, 2002).

263. Schemo, *supra* note 263.

264. Letter from T.P. Tumelty, Head of Candidate Guidance at the Naval Academy to Daniel A. Wurangian (Nov. 27, 2001) (on file with author). I obtained a copy of this letter from Christina Perez of FairTest in Cambridge, Massachusetts.

265. *Id.*

266. *See* Letter from Christina Perez of FairTest, *supra* note 263.

267. *See* Jeffrey Selingo, *Civil-Rights Groups Blast Florida's Use of SAT Scores in Awarding Scholarships*, CHRON. HIGHER EDUC., Dec. 21, 2001, at A18 (also reporting that African Americans comprise 14.4% of Florida SAT test-takers, but only 3% of Bright Futures level one scholarship recipients).

of African Americans received these awards.[268]

A university defending its use of the SAT in admissions would undoubtedly rely upon the large body of studies produced by ETS and College Board researchers purporting to validate the SAT as a predictor of first year college grades.[269] This position hardly demonstrates educational necessity, however. Scholarly critics of the SAT, some of whom might be retained as expert witnesses by plaintiffs counsel in educational disparate impact litigation have for many years pointed out that combining the SAT with high school grades only incrementally improves the prediction of freshman grade point average compared to high school grades (HSGPA) alone.[270] For example, researchers at the University of California Office of the President recently completed a study of 78,000 freshmen who entered seven UC campuses between 1996 and 1999.[271] The authors found that HSGPA explained 15.4% of the variance in freshman grades among enrolled students at UC campuses, HSGPA combined with SAT scores explained 20.8%, HSGPA combined with the SAT II subject-specific achievement tests explained 22.2%, and HSGPA, SAT, and SAT II combined explained 22.3%.[272] Based on these results, if UC were sued over the disparate impact of the SAT, it would be difficult to advance "a substantial legitimate justification" for the SAT when the test improves the percentage of variance explained by a statistically insignificant 0.1% above that explained by HSGPA and the SAT II, and when the SAT only adds 5.4% to the variance explained by HSGPA alone. At UC Berkeley, UCLA, and UC San Diego—the most selective campuses,

268. *See* Press Release, MALDEF/FairTest, Florida State Scholarship Program Unfairly Discriminates, Say Civil Rights and Educational Groups (Aug. 26, 2002) (listing Academic Scholars Awards for 1999-2001).

269. *See, e.g.*, Brent Bridgeman et al., *Predictions of Freshman Grade-Point Average from the Revised and Recentered SAT I: Reasoning Test* (2000), COLLEGE BOARD RESEARCH REPORT NO. 2000-1; WARREN W. WILLINGHAM ET AL., PREDICTING COLLEGE GRADES: AN ANALYSIS OF INSTITUTIONAL TRENDS OVER TWO DECADES (1990); Rick Morgan, *Predictive Validity Within Categorizations of College Students: 1978, 1981, and 1985* (1990), ETS RESEARCH REPORT NO. 90-14; Rick Morgan, *Analyses of the Predictive Validity of the SAT and High School Grades From 1976 to 1985* (1989), COLLEGE BOARD RESEARCH REPORT NO. 89-7.

270. *See, e.g.*, CROUSE & TRUSHEIM, *supra* note 249, at 52; James Crouse, *This Time the College Board Is Wrong*, 55 HARV. EDUC. REV. 478, 479 (1985); Peter Sacks, *Standardized Testing: Meritocracy's Crooked Yardstick*, CHANGE, Mar./Apr. 1997, at 25-26; Warner V. Slack & Douglas Porter, *The Scholastic Aptitude Test: A Critical Appraisal*, 50 HARV. EDUC. REV. 154, 165 (1980).

271. *See, e.g.*, SAUL GEISER & ROGER STUDLEY, UC AND THE SAT: PREDICTIVE VALIDITY AND DIFFERENTIAL IMPACT OF THE SAT I AND SAT II AT THE UNIVERSITY OF CALIFORNIA (2001). UC Santa Cruz was excluded because in many courses that institution issued narrative evaluations rather than letter grades. *See id.*

272. *See id.* at 3 tbl.1.

and the ones most likely to be named as defendants in a disparate impact lawsuit—the relative contribution of the SAT was even lower than it was for the UC system overall.[273] To summarize, the SAT imposes a substantial adverse effect on underrepresented minority students, yet its contribution to the prediction of freshman grades is quite modest. Moreover, the UC system's educational necessity position was most likely undermined when UC President Richard Atkinson, who himself is a cognitive psychologist steeped in the testing literature, suggested that the UC system could discontinue using the SAT in favor of another test that it might develop.[274]

More importantly, the ability to predict freshman grades in college is hardly dispositive for a defendant university attempting to meet its educational necessity burden in a Title VI disparate impact claim. A strong argument can be made that college graduation is of greater ultimate importance than freshman GPA, and the educational necessity of the SAT is even more questionable considering available data on graduation patterns. U.S. Department of Education research analyst Clifford Adelman argues:

> The justification for using SAT scores in admission decisions is that they are a decent predictor of first-year college grades. True, but so what? That criterion has nothing to do with the principal goal of students at four-year colleges and their families: completing a bachelor's degree. Nor do state legislatures give a hoot about grades when they judge the performance of public universities: Performance means *graduation rates*.[275]

Using the U.S. Department of Education's comprehensive national database, Adelman found that, after controlling for major background characteristics of students, the quality and intensity of high school academic curriculum was a far better predictor of degree completion than SAT scores.[276] Another major national study by

---

273. *See id.* at 5-6. Note that the UC finding is not a byproduct of restriction of range. The range of student SAT scores would be expected to be more restricted at the most competitive UCs. However, restriction of range would not explain the predictive inferiority of the SAT I in comparison to the SAT II, since the variances of SAT I and SAT II score are quite similar within each school. For the same reason, restriction of range cannot explain why the inferiority of SAT I is greater at UCB, UCLA, and UCSD than other UCs. *See id.* at 4 n.8.

274. *See supra* Part I.A.

275. Clifford Adelman, *Why Can't We Stop Talking About the SAT?*, CHRON. HIGHER EDUC., Nov. 5, 1999, at B4.

276. *See* Clifford Adelman, *Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degee Attainment* (1999), *at* http://www.ed.gov/pubs/Toolbox/

UCLA Professor Alexander Astin looked at longitudinal data from the Cooperative Institutional Research Program (CIRP), and found that the SAT only correlated 0.27 with graduation rates, meaning that the SAT only explained 5% of the variance in graduation rates.[277]

Do SAT scores have a stronger association with graduation rates at highly selective universities, which are by and large the institutions at issue in the affirmative action debate? According to Abigail and Stephan Thernstrom, prominent critics of affirmative action, UC Berkeley graduation rates "correlated perfectly with SAT scores."[278]

(last visited June 14, 2002); Clifford Adelman, *The Rest of the River*, UNIV. BUS., Jan.-Feb. 1999, at 42, 48.

277. *See* ALEXANDER W. ASTIN, WHAT MATTERS IN COLLEGE 193 (1993) (reporting for a sample of 38,000).

278. Abigail Thernstrom & Stephan Thernstrom, Letter to the Editor, N.Y. TIMES, June 1, 1998. This argument is laid out in greater detail in THERNSTROM & THERNSTROM, *supra* note 44, at 406-07. In particular, the Thernstroms argue that SAT-band data from Berkeley's 1988 entering class (reproduced in the left columns of the table below) establishes that the SAT correlates strongly with graduation rates. *See id.* at 407 tbl.8.

However, Gregg Thomson, Director of the Office of Student Research at UC Berkeley, offers what we believe is a persuasive rebuttal to the Thernstroms' claim. Gregg Thomson, Is the SAT a "Good Predictor" of Graduation Rates? The Failure of "Common Sense" and Conventional Expertise and a New Approach to the Question (1998) (unpublished paper presented at the California Association of Institutional Research annual meeting). Thomson argues that the Thernstroms' data presentation is misleading because the cells with far lower graduation rates (SATs in the 700s and 800s) only include 2% and 4% of the cohort, respectively. *See id.* at 4-5. As indicated in the far right column in the table below, after recalculating admission rate averages for nine equally sized intervals, the SAT-graduation rate association diminishes considerably. As indicated by the middle-right column below, even within the Thernstroms' reporting format, the SAT-graduation rate correlation decreases considerably after taking out students admitted by exception, which is a classification (distinct from affirmative action) for those who did not meet the basic UC eligibility requirements, which largely includes recruited athletes. *See id.* at 5. Finally, Thomson reports that there is "zero correlation" between SAT scores and eventual graduation rates for the African Americans within the same cohort of Berkeley students discussed by the Thernstroms. *See id.* at 6.

Thernstroms' Data on 1988 Berkeley Freshmen Who Graduated Within Six Years (at 407 tbl.8)		Thomson's Data on 1988 Berkeley Freshmen Who Graduated Within Six Years (at 4-5)	
SAT BAND	Graduation Rate	Graduation Rates by SAT Band without "admissions by exception" (mostly recruited athletes)	Graduation Rates by SAT After Dividing Berkeley's Entering Class into Nine SAT intervals With Equal Numbers of Students
700-799	58%	73%	77%
800-899	62%	75%	80%
900-999	72%	79%	86%

However, more reliable information is presented in The Shape of the River, in which William Bowen and Derek Bok, the former presidents of Princeton and Harvard, respectively, extensively studied the College and Beyond (C&B) database of twenty-eight (mostly private) elite colleges and universities.[279] Bowen and Bok found that, after controlling for school selectivity, high school grades, socioeconomic status, and other characteristics, SAT scores bore little relationship to graduation rates (and no relationship above scores of 1000). Students with SAT scores under 1000 had graduation rates of 83%, students in the 1000s had rates of 86%, those in the 1100s had rates of 88%, those in the 1200s had rates of 86%, and those above 1300 graduated 87% of the time.[280] The SAT findings contrasted with school selectivity, which continued to be associated with graduation rates after controlling for other factors.[281] Bowen and Bok also report that at the most selective C&B schools African Americans with SAT scores under 1000 had graduation rates of 88%, whereas in the least selective C&B schools even African American students with SAT scores over 1300 had graduation rates of 75%.[282] In summary, the C&B data suggests that factors other than the SAT—those having to do with institutional resources (endowment size, class size, the availability of support programs, etc.)—are much more influential determinants of graduation rates.

The Bowen and Bok data should also remind researchers that

1000-1099	78%	82%	88%
1100-1199	83%	86%	86%
1200-1299	86%	87%	89%
1300-1399	88%	91%	92%
1400-1499	84%	86%	88%
1500-1599	79%	82%	86%

We point out these differences in data presentation and interpretation because much of the public debate about affirmative action, merit, and the SAT involves UC Berkeley.

279. *See* WILLIAM G. BOWEN & DEREK BOK, THE SHAPE OF THE RIVER (1998). Seventy percent of C&B students attended private colleges and universities, while 30% attended four large public universities. *See id.* at xxxvii.

280. *See id.* at 66 fig.3.6.

281. *See id.* at 63. Bowen and Bok conclude:

> The central finding is that the effect of school selectivity on graduation rates persists after controlling not only for differences in SAT scores, but for other factors as well. In other words, among students of the same gender with similar SAT scores, high school grades, and socioeconomic status, those who attended the most selective schools graduated at higher rates than did those who attended less selective schools.

Id.

282. *See id.* at 61 fig.3.3.

much of the SAT-graduation rate correlation reported in other large-scale studies may be an artifact of combining data from different schools, while failing to acknowledge that the most well-endowed elite private schools simultaneously have greater institutional resources, as well as higher graduation rates and higher SAT scores.[283] For example, a recent study by Burton and Ramist of the College Board summarized eight studies of SAT and graduation rates, and reported a 0.33 correlation between these two measures.[284] Yet, Burton and Ramist acknowledge that their estimate may be too high because they could not account for institutional effects.[285]

To give a practical example, this means that it would be incorrect to combine data from Harvard and California State University at Hayward, and then draw conclusions about the SAT's ability to forecast graduation rates without first controlling for institutional and student background characteristics. First, graduation rates for Harvard will reflect the benefits of receiving an education at a place with enormous institutional resources (students receive greater individualized attention from faculty and administrators, stronger peer support networks, etc.). Second, Cal State Hayward students are far more likely to encounter socio-economic barriers that make uninterrupted graduation more difficult for reasons unrelated to academic preparation or ability. The SAT's correlation with income and other measures of socioeconomic status[286] and institutional

283. *See* ZWICK, *supra* note 63, at 93-94. In a review of the literature on standardized tests and graduation rates, Zwick cautions:

> In a large study that includes many colleges, there will be a much larger range of test scores and graduation rates than in a single school. Multi-institution analyses of graduation are usually based on the combined data from all the schools (unlike multi-institution GPA prediction studies, which usually involve analyses that have been conducted *within* institutions and then averaged). To some extent, then, the apparent association between test scores and graduation will reflect the fact that some *schools* have both higher test scores and higher gradation rates than others.

Id.

284. *See* Nancy W. Burton & Leonard Ramist, *Predicting Success in College: SAT Studies of Classes Graduating Since 1980*, at 16 tbl.9 (2001), COLLEGE BOARD RESEARCH REPORT NO. 2001-2.

285. *See id.* at 17 ("Most of the results in Table 9 are based on multi-institution studies, so the tendency of more selective institutions to have higher graduation rates will affect the correlations. Pending further research, one cannot be sure what part of a correlation in Table 9 is due to the institution-level relationship of selectivity to retention and what part is due to the predictability of individual students' graduation from their grades and SAT scores.").

286. The relationship between SAT scores and income is indicated in the table below, which is based on test-takers' self-reported parental income for all high school seniors who took the SAT in 2001. *See* FairTest, *University Testing: 2001 SAT Scores*, *at* http://www.fairtest.org/univ/2001SAT%20Scores.html (last visited June 28, 2002).

resources[287] tends to artificially boost the correlation between graduation rates and SAT scores when combining Harvard and Hayward data. Consequently, when Warren Willingham of ETS studied SAT-graduation relationships *within* each of nine colleges, the correlation coefficient dropped to only 0.15.[288]

In analyzing whether "a substantial legitimate justification" exists for over-reliance on the SAT despite its disparate impact, a key consideration is that there must be a fit between a university's mission and its admission practices. In a recent report on standardized tests, the prestigious National Academy of Sciences recommended that "[a]dmissions policies and practices should be derived from and clearly linked to an institution's overarching intellectual and other goals" and that the "use of test scores in the admissions process should serve those institutional goals."[289] While these recommendations may seem like common sense, universities espousing lofty institutional missions frequently fail to carefully consider whether or not their admission criteria are well-suited to serve important goals.[290] For example, in its

Family Income	Combined SAT Scores
Under $10,000	864
$10,000 - $20,000	898
$20,000 - $30,000	942
$30,000 - $40,000	976
$40,000 - $50,000	1004
$50,000 - $60,000	1011
$60,000 - $70,000	1035
$70,000 - $80,000	1049
$80,000 - $100,000	1074
$100,000+	1126

For further discussion of the relationship between SAT scores and income level, see Sturm & Guinier, *supra* note 140, at 970, Sacks, *supra* note 271, at 25-26. To be clear, we are not claiming that the correlation between SAT scores and income is entirely a reflection of bias in the SAT. The unfortunate fact is that since poor and affluent students have unequal educational opportunities, income-based differences in SAT scores are hardly surprising for reasons unrelated to test bias. We are making the more technical point that it is questionable logic to assume that the correlation between SAT scores and graduation rates is *caused* by SAT-related skill differences without first accounting for other factors (socioeconomic status, institutional effects, etc.) that correlate with SAT scores.

287. *See* BOWEN & BOK, *supra* note 280.

288. *See* Burton & Ramist, *supra* note 285, at 17 (citing WARREN W. WILLINGHAM, SUCCESS IN COLLEGE: THE ROLE OF PERSONAL QUALITIES AND ACADEMIC ABILITY (1985)).

289. Alexandra Beatty et al., *Myths and Tradeoffs: The Role of Tests in Undergraduate Admissions* (1999), *at* http://www.nap.edu/html/myths_tradeoffs/#Summary (last visited June 14, 2002).

290. Lani Guinier, *Confirmative Action*, 25 LAW & SOC. INQUIRY 565, 578 (2000) ("Law schools, especially public institutions like the University of Michigan, could at least

answer to the *Rios/Castaneda* complaint filed by civil rights organizations, defense counsel for UC Berkeley stated that Berkeley's institutional mission was to "admit students who, among other characteristics, demonstrate exceptional achievement and talent, who will contribute to the campus community, and will bring diversity of personal experience and background."[291] The SAT's relationship to such criteria is far from self-evident. For example, a thirty-year retrospective study of three classes of Harvard University alumni found that *low* SAT scores and a blue-collar background correlated with measures of success such as community involvement, professional satisfaction, and high income.[292]

Evidence suggests that the SAT and other standardized tests are particularly weak predictors of potential contributions to community service and similar "public spirited" institutional goals. For example, Bowen and Bok found that within the C&B database, African American graduates, many of whom received affirmative action consideration, and who had average SAT scores over 200 points lower than Whites, were nonetheless significantly more likely than their White classmates to become the leaders of civic service organizations,

---

be more explicit and more open about their real mission, and express a willingness to abandon those rigid entry-level criteria that do not predict the kinds of behavior among their graduates that the school purports to value."); Thomas D. Russell, *The Shape of the Michigan River as Viewed from the Land of* Sweatt v. Painter *and* Hopwood, 25 LAW & SOC. INQUIRY 507, 511 (2000) ("As part of the defense of race-conscious affirmative action at state universities like Michigan and UT, the faculty and administrators, as well as their lawyers, ought to think hard about the aims of the universities in light of their character as *state institutions*."); Note, *The Relationship Between Equality and Access in Law School Admissions*, 113 HARV. L. REV. 1449 (2000). The author of this note observes:

> [T]he institution must define merit in a way that enables the institution to create selection criteria that evaluate the skills necessary for participation within the institution. If the selection criteria identify and reward other attributes, access is granted arbitrarily because individuals are chosen based on something other than their capacity to engage in the activity at issue. Such a procedure not only prevents institutions from selecting the best candidates, but it can also have an unnecessary discriminatory effect on certain groups. Despite these potential problems, institutions rarely examine or reform their selection criteria to ensure that the criteria accurately identify individuals who will enable the institution to accomplish successfully its mission.

Id. at 1456.

291. Rios v. Regents of the University of California, Answer to First Amended Complaint at 9, April 9, 1999 (N.D. Cal., Case No. C 99-0525 SI).

292. Sturm & Guinier, *supra* note 140, at 976-77 (citing David K. Shipler, *My Equal Opportunity, Your Free Lunch*, N.Y. TIMES, Mar. 5 1995, section 4 at 1, 16). Admittedly, Harvard has one of the most competitive applicant pools in the country, so restriction of range effects caution against over-interpretation. On the other hand, there is reason to think that similar results might obtain at other elite universities and colleges.

including those in law, medicine, business, and other professions.[293] A study of alumni of the University of Michigan Law School graduating classes of 1970-1996 found similar results.[294] Moreover, this is not simply a self-selection effect of admission policies at these institutions, as other research indicates that within nationally representative samples of applicants, standardized tests such as the LSAT, GRE, and MCAT negatively correlate with valuing social activism, leadership, and concern for others.[295] Some institutions, such as Bates College, actually find that making the SAT optional allows them to better fulfill their institutional mission, and has the added bonus of broadening and deepening their applicant pool.[296]

3. *Evaluating Equally Effective but Less Discriminatory Alternatives*

Plaintiffs may prevail in a Title VI disparate impact lawsuit even after the defendant provides sufficient evidence of educational necessity if plaintiffs can meet their burden, and demonstrate that an alternative practice results in smaller racial/ethnic disparities but is nonetheless equally effective in meeting the institution's educational goals.[297] The courts can consider the administrative feasibility of suggested alternatives, including differences in cost and time.[298]

---

293. *See* BOWEN & BOK, *supra* note 280, at 29-31, 160-68.

294. *See* Richard O. Lempert et al., *Michigan's Minority Graduates in Practice: The River Runs Through Law School*, 25 LAW & SOC. INQUIRY 395, 485-90 (2000).

295. *See* Kidder, *supra* note 15, at 55-56. *See also* Astin, *supra* note 278, at 202-09, 213; Leonard L. Baird, *Biographical and Educational Correlates of Graduate and Professional School Admission Test Scores*, 36 EDUC. & PSYCHOL. MEASUREMENT 415, 418-19 (1976).

296. *See* William C. Hiss, *Optional SAT's at Bates: 17 Years and Not Counting*, CHRON. HIGHER EDUC., Oct. 26, 2001, at B10 (also noting that the Bates students who do not submit their SAT scores have GPAs and graduation rates equal to students who do submit SAT scores).

297. *See* U.S. DEP'T OF EDUC. OFFICE FOR CIVIL RIGHTS, *supra* note 92, at 58.

298. *See id.* at 59 n. 203. *See also* Sharif v. New York State Educ. Dep't, 709 F. Supp. 345, 363-64 (S.D.N.Y. 1989) (rejecting New York's argument that alternatives to sole reliance on the SAT in awarding scholarships were impractical in light of the fact that several other states employed alternative criteria which resulted in smaller gender disparities); GI Forum v. Texas Educ. Agency, 87 F. Supp. 2d 667, 682 (W.D. Tex. 2000) (ruling, in the context of a state standardized test required to graduate from high school, "[t]he Plaintiffs produced no alternative that adequately addressed the goal of systemic accountability"); Cureton v. Nat'l Collegiate Athletic Ass'n, 37 F. Supp. 2d 687, 714, *rev'd on other grounds*, 198 F.3d 107 (3d Cir. 1999) ("Plaintiffs have shown at least three alternative practices resulting in less racial disproportionality while still serving the NCAA's goal of raising student-athlete graduation rates . . . That is all the proof that Plaintiffs need to demonstrate under Title VI.").

We wish to make clear at the outset that establishing the existence of equally effective but less discriminatory alternatives in a Title VI disparate impact case is quite distinct from the narrow tailoring prong of strict scrutiny review in Equal Protection challenges to university affirmative action programs.[299] Thus, while we discuss percentage plans in this portion of the article, we warn readers not to mistakenly interpret our analysis to mean that race-conscious admission programs at institutions such as the University of Michigan[300] and the University of Georgia[301] are not narrowly tailored to advance a compelling governmental interest. We also wish to avoid conveying the impression that pervasive inequalities in K-12 education are excused by virtue of percentage plans; it is only that K-12 equity issues are beyond the scope of this article.[302]

One important source of data on equally effective but less discriminatory alternatives to the SAT is the Texas "Ten Percent Plan," which was enacted by the Texas legislature and signed by then-Governor George W. Bush in 1997, shortly after the Fifth Circuit's *Hopwood v. Texas* ruling banned affirmative action.[303] The Ten Percent Plan allows students graduating in the top tenth of their high school class to gain admission to the University of Texas-Austin (UT-Austin), Texas A&M University, and other campuses, without regard to performance on the SAT. In Table 1, we display UT-Austin freshman enrollment data by race/ethnicity for the five years since affirmative action was prohibited. The 1997 figures were after *Hopwood* banned race-sensitive admissions, but were before the Ten Percent Plan took effect. The pre-Ten Percent Plan numbers for the 1997 class therefore provide a useful baseline to compare with the subsequent four classes admitted under this plan. The data indicate that the proportion of African Americans and Latinos at UT-Austin have improved modestly (and slightly more for APAs) after the plan took effect. African Americans were 2.7% of freshman enrollments in 1997, compared to an average of 3.6% during 1998-2001. Latinos made up 12.6% of UT-Austin freshman enrollments in 1997, compared

299. For a summary of several recent educational affirmative action cases involving narrow tailoring, see Kidder, *supra* note 25, at 179, 193, 202-04. For in-depth discussion of the issue, see Ian Ayres, *Narrow Tailoring*, 43 UCLA L. REV. 1781 (1996).

300. *See* Gratz v. Bollinger, 122 F. Supp. 2d 811 (E.D. Mich. 2000).

301. Johnson v. Bd. of Regents of the Univ. System of Georgia, 263 F.3d 1234 (11th Cir. 2001).

302. *See infra* Part V.

303. *See* Holley & Spencer, *supra* note 32, at 252-59.

to an average of 13.5% during 1998-2001. While the parties would likely dispute causation, this kind of data should be sufficient to make a showing that percentage plans can be a less discriminatory alternative to post-affirmative action admissions in which the SAT is required.[304]

TABLE 1:
Post-*Hopwood* Freshman Enrollments at UT-Austin 1997-2001[305]

	1997	**1998**	**1999**	**2000**	**2001**
Black	190 (2.7%)	199 (3.0%)	286 (4.1%)	296 (3.9%)	242 (3.3%)
Latino	892 (12.6%)	891 (13.2%)	976 (13.9%)	1011 (13.2%)	1024 (14.0%)
APA	1130 (15.9%)	1133 (16.8%)	1221 (17.3%)	1325 (17.2%)	1413 (19.2%)
White	4730 (66.8%)	4399 (65.2%)	4447 (63.2%)	4801 (62.5%)	4447 (60.6%)

The remaining issue is whether a policy like the Texas Ten Percent Plan can be an equally effective alterative to reliance on the SAT. Again, data from the flagship UT-Austin campus are illuminating. Students in the top 10% of their high school class earned

304. One study by David Montejano analyzed UT-Austin's feeder high schools, and found that the principal beneficiaries of the Ten Percent Plan were Black and Chicano students from inner-city high schools in San Antonio, Houston, and Dallas, as well as Whites from rural high schools in northern and eastern Texas. *See* David Montejano, *Access to the University of Texas at Austin and the Ten Percent Plan: A Three-year Assessment* (2001), *at* http://www.utexas.edu/student/research/reports/admissions/Montejanopaper.htm (last visited June 14, 2002). *See also* David Montejano, *Maintaining Diversity at the University of Texas, in* RACE AND REPRESENTATION: AFFIRMATIVE ACTION 359 (Robert Post & Michael Rogin eds., 1998).

In a SAT disparate impact case, plaintiffs' and defendants' experts could be expected to dispute how much improvement in racial/ethnic composition is attributable to not considering the SAT, as opposed to shifting demographics of the applicant pool, increased recruiting efforts, changes in financial aid availability, etc.

305. The information in Table 1 was generated from several different sources. *See* UT Austin Office of Institutional Studies' Enrollment Tables (2000); GARY M. LAVERGNE & BRUCE WALKER, IMPLEMENTATION AND RESULTS OF THE TEXAS AUTOMATIC ADMISSIONS LAW (HB 588) AT THE UNIVERSITY OF TEXAS AT AUSTIN REPORT NUMBER 4 (2001), *available at* http://www.utexas.edu/student/research/reports/admissions/HB588-Report4.pdf (last visited June 14, 2002); Holley & Spencer, *supra* note 32, at 252 tbl.1. We did not include American Indians in Table 1 because their numbers at UT-Austin, ranging from twenty-eight to thirty-seven annually, were too small to form the basis of conclusions vis-à-vis the Ten Percent Plan.

significantly higher freshman grades than non-top 10% students—this finding was true overall for each racial and ethnic group, and within each field of study.[306] In fact, top 10% students with SAT scores in the 1200s had higher freshman GPAs than non-top 10% students with SATs in the 1400-1600 range, top 10% students with SAT scores in the 1000s had higher GPAs than non-top 10% students with SATs in the 1200s, and so forth.[307] Persistence and graduation rates were likewise higher at UT-Austin for top 10% students than those not in the top 10%.[308]

Some readers may reasonably find that the above performance data on the Texas Ten Percent plan is not an entirely satisfactory comparison, since many of the students in the top 10% of their high school class also had high SAT scores and would have been admitted to UT-Austin regardless of the Ten Percent Plan. Similarly, several scholars have criticized other major affirmative action studies for not separating students of color who would have been admitted anyway from those who would not have been admitted but for affirmative action.[309] However, Bowen and Bok and other researchers have

306. *See* LAVERGNE & WALKER, *supra* note 306, at 7-13.

307. *See id.* at 7 tbl.VI.

308. *See id.* at 16-20. *See also* Montejano, *Maintaining Diversity*, *supra* note 305, at 2 (noting that top 10% students have outperformed non-top 10% students with SAT scores 200-300 points higher). Another study of public university students in Indiana likewise found that if students' SAT scores were subtracted from the mean SAT scores for their high schools (which is, in effect, similar to the Texas Ten Percent Plan) the "merit-aware" index scores were equally effective as predictors of student persistence compared to unadjusted SAT scores. *See, e.g.*, Edward P. St. John, *Aptitude vs. Merit: What Matters is Persistence*, 24 REV. HIGHER EDUC. 131 (2001).

309. Terrence Sandalow, *Rejoinder*, 97 MICH. L. REV. 1923 (1999). Sandalow criticizes Bowen and Bok:

> In The Shape of the River, presidents Bowen and Bok pronounce the race-sensitive admission policies adopted by selective undergraduate schools a resounding success. The evidence they adduce in support of that conclusion primarily concerns the performance of African-American students in and after college. But not all African-American students in those institutions were admitted in consequence of minority preference policies. Some, perhaps many, would have been admitted under race-neutral policies. I argued at several points in my review that since these students might be expected to be academically more successful than those admitted because of their race, the evidence on which Bowen and Bok rely provides a potentially distorted view of the latter's performance, almost certainly suggesting a greater level of success than those students actually achieved.

Id. at 1923. *See also* Terrance Sandalow, *Minority Preferences Reconsidered*, 97 MICH. L. REV. 1874 (1999). Richard Sander, *The Tributaries to the River*, 25 LAW & SOC. INQUIRY, 557, 559 n.2 (2000) (In criticizing Lempert, Chambers, and Adams study of the University of Michigan Law School, Sander argues: "It is worth pointing out that in all the paper's analyses, 'minority' is implicitly used as a proxy for 'affirmative action admit.' Given the extent of background information the authors had, I suspect they could have identified

correctly noted that framing the debate in this manner is to chase an impossible goal, because it is surprisingly difficult to know as an empirical matter which students of color were and were not admitted under affirmative action.[310] Accordingly, we approached UT-Austin

---

which students were in fact probably admitted through affirmative action, and which students would have been admitted through a race-blind process. This would have made more convincing those analyses that purport to assess the effects of affirmative action."); Robert L. Nelson & Monique Payne, *Minority Graduates from Michigan Law School: Differently Successful*, 25 LAW & SOC. INQUIRY 521, 522 (2000) ("Minority status is then an imperfect indicator of whether an applicant was admitted preferentially on the basis of race. In an article primarily concerned with assessing the effects of affirmative action policies, blurring the distinction between minority and preferential admissions is problematic because it may obscure some fundamental differences within the group labeled minority. For example, perhaps those minorities who were admitted without preferential treatment were more likely to succeed than others granted admission.").

310. *See* William G. Bowen & Derek Bok, *Response to Review by Terrance Sandalow*, 97 MICH. L. REV. 1917 (1999). Bowen and Bok observe:

> There is absolutely no way of knowing when race was and was not dispositive (or, to put the question another way, which African-American candidates would have been admitted had they been White). And, in fact, even framing the question this way is to chase a will o' the wisp. As one admissions dean put it in a recent conversation, people have to understand that we look at all the attributes of a candidate together; we view the race of a candidate in conjunction with so many other things—what school the student attended, where and how he or she grew up, leadership potential, 'drive,' and so on. Moreover, in deciding whether or not to admit a particular candidate, we also consider who else has already been admitted to the class. This admissions officer went on to say that, even with all the information he has (far more than would ever be available to any outside student of the process), he himself could not say which candidates were and were not admitted solely because of their race.

Id. at 1918-19. *See also* Richard O. Lempert et al., *Michigan's Minority Graduates in Practice: Answers to Methodological Queries*, 25 LAW & SOC. INQUIRY 585 (2000).

Responding to criticism of their study (cited in the previous footnote) Lempert, Chambers, and Adams argue:

> Nelson and Payne and Sander would all like to know what our results would look like if we had excluded from our minority sample minority graduates who would have been admitted to Michigan without a boost from affirmative action. Their concern is that the success of these graduates explains why minority status and admissions credentials seem not to explain current income or career satisfaction. We understand why they are curious and concerned, but there is a good argument that the groups should not be separated. The success of minorities who would have been admitted to Michigan without affirmative action may be due in considerable measure to the existence of the program . . . . Moreover, if we turn from theory to practice, it is impossible to identify with certainty most of those minority students at Michigan who would have been admitted had the school not had an affirmative action program. Many minority students with admissions indexes in the range of White admittees nevertheless benefited at the admissions stage from Michigan's affirmative action program. This is because, like most of their white counterparts, most minority students with admissions indexes sufficient for admission to Michigan without affirmative action nonetheless do not have quantitative credentials so strong . . . . Michigan's concern for diversity meant that all these students presented very strong cases for admission, and we have no way of

officials about obtaining more accurate data on this point, but they were unable to provide it for the same reason.[311]

E. *The Viability of Filing Complaints with the Department of Education*

The U.S. Department of Education regulations interpreting Title VI prohibit educational institutions that receive federal funding from using criteria (in admissions, scholarship allocation, etc.) that have an unwarranted disparate impact on students of color.[312] In addition to using section 1983 as a mechanism to privately enforce the Department of Education's disparate impact regulations, the costs and benefits of filing a complaint directly with Office for Civil Rights (OCR) should also be explored below. A recent example is the OCR complaint filed by MALDEF, FairTest, and other organizations over Florida's use of a 1270 SAT cutoff score for the state's $164 million "Bright Futures" scholarship program.[313]

For many public interest organizations constrained by the cost of litigation, the lower cost of filing an OCR complaint may be more appealing, even though there are serious drawbacks. One glaring limitation is that a complainant does not possess a right to participate in an OCR investigation.[314]

From the plaintiffs' perspective, the built-in level of passivity in an OCR investigation is substantial, which makes it difficult to use such a complaint as a lightening rod for the larger political movement for educational equity. Consequently, we conclude that an OCR complaint will usually fail what might be called the "social justice praxis test," although litigation often fails this test as well. For instance, environmental law Professor Luke Cole advocates "practicing

---

distinguishing most of those students who would have made it had a concern for diversity not existed from those who would not have been admitted.

Bowen & Bok, *supra* note 311, at 593-94.

311. Specifically, we contacted Gary Lavergne, Director of Admissions Research at UT-Austin, and author of several reports on the Texas Ten Percent Plan. We also requested data from Professor Leicht at the University of Iowa, who heads a Ford Foundation study of the Texas Ten Percent Plan. Lavergne could not provide us with the data for reasons similar to those cited by Bowen and Bok, Lempert, and Chambers and Adams.

312. *See* 34 C.F.R. § 100.3(vii)(2) (Lexis 2002).

313. See Andrea Robinson, *Coalition Alleging Bias in Fla. Scholarship Program*, MIAMI HERALD, Aug. 27, 2002, at A1.

314. *See* Mank, *supra* note 189, at 363 (noting that Title VI administrative investigations do not protect the individual rights of the complainant).

law in a way that empowers people, that encourages the formation and strengthening of client groups, and that sees legal tactics in the context of broader [political] strategies."[315] Civil rights scholar Eric Yamamoto espouses a similar notion of "critical race praxis," which involves using the courts as part of a larger communicative process "to help focus cultural issues, to illuminate institutional power arrangements, and to tell counter-stories in ways that assist in the reconstruction of intergroup relationships and aid larger social-political movements."[316]

In summary, we do not mean to disparage those who decide to file OCR complaints in order to enforce Title VI disparate impact regulations. Indeed, our analysis of the case law indicates that OCR complaints may be the only viable legal remedy in those jurisdictions that no longer recognize a private right of action to enforce Title VI disparate impact regulations. Rather, we emphasize the need to think strategically about how filing an OCR complaint (as well as filing a lawsuit) can contribute to a larger movement to advance educational equity issues.

V. CONCLUSION

According to the 2000 Census, nearly thirty-three million Latinos, including twenty-two million Chicanos, live in the United States.[317] More than half of U.S. Latinos reside in California and Texas,[318] where Proposition 209 and *Hopwood* currently prohibit the consideration of race in public university admissions. Consequently, while Latinos comprised 32.5% of Californians in the 2000 Census (and more than a third of California's public high school graduates in 2002),[319] Latinos comprised 12.7% of freshmen enrollments at all UC campuses in the

315. Luke Cole, *Empowerment as the Key to Environmental Protection: The Need for Environmental Poverty Law*, 19 ECOLOGY L.Q. 619, 648 (1992). *See also* GERALD P. LOPEZ, REBELLIOUS LAWYERING (1992).

316. Eric K. Yamamoto, *Critical Race Praxis: Race Theory and Political Lawyering Practice in Post-Civil Rights America*, 95 MICH. L. REV. 821, 885-86 (1997). *See also* ERIC K. YAMAMOTO, INTERRACIAL JUSTICE (1999).

317. *See* Melissa Therrien & Roberto R. Ramirez, *The Hispanic Population in the United States: March 2000* (2001), *in* U.S. CENSUS BUREAU REPORT, *available at* http://www.census.gov/population/www/cen2000/briefs.html (last visited June 25, 2002).

318. *See* Press Release, Census 2000, U.S. Census Bureau (May 2001), *available at* http://www.census.gov/press-release/www/2001/cb01-81.html (last visited June 25, 2002).

319. *See* Bob Laird, *Bending Admissions to Political Ends*, CHRON. HIGHER EDUC., May 17, 2002, at B11 (UC Berkeley's former Director of Admissions, citing data from the California Department of Finance).

first four years following the ban on affirmative action (1998-2001).[320] Latino representation in 1998-2001 was even lower at UC Berkeley (9.6%), and at UC San Diego (8.9%), another highly selective campus where admission is driven by SAT scores and grades to an even greater extent than at Berkeley.[321] Likewise, while Latinos comprised 32% of Texas residents in the 2000 Census, they made up 13.4% of freshman enrollments at the University of Texas-Austin in the five years following the *Hopwood* decision (1997-2001).[322]

Post-affirmative action university admission data are even more discouraging for African Americans, who comprised just under 3% of 1998-2001 freshmen enrollments in the UC system.[323] According to the 2000 Census, African Americans comprised 29.2% of Georgia residents.[324] However, at the University of Georgia, under the quite modest 1999 affirmative action plan that was recently struck down by the Eleventh Circuit,[325] African Americans comprised less than 6% of freshmen enrollments.[326] This reflected the reality that approximately 85% of freshmen at the University of Georgia (which had a 160-year history of de jure segregation) were admitted solely of the basis of SAT/GPA index scores, and that within the smaller pool of students receiving comprehensive review, the plus factor given to race was less

320. *See* University of California Office of the President, *Application, Admissions and Enrollment of California Resident Freshmen for Fall 1995 through 2001* [hereinafter California Resident Freshmen] *at* http://www.ucop.edu/news/factsheets/flowfrc9501.pdf (last visited June 25, 2002). For more extensive policy discussion of Latino's lack of access to higher education in California, see generally Richard Delgado & Jean Stefancic, *California's Racial History and Constitutional Rationales for Race-Conscious Decision Making in Higher Education*, 47 UCLA L. REV. 1521 (2000); Jorge H. del Pinal, *Latinos and California's Future: Too Few at the School's Door*, 10 LA RAZA L.J. 631 (1998); Aida Hurtado et al., *Becoming the Mainstream: Merit, Changing Demographics, and Higher Education in California*, 10 LA RAZA L.J. 645 (1998); Rachel F. Moran, *Unrepresented*, 55 REPRESENTATIONS 139 (1996).

321. *See* California Resident Freshmen, *supra* note 321 (listing enrollments by campus and race/ethnicity); REBECCA ZWICK, *supra* note 63, at 38 (describing UC San Diego's 1999 admissions policy based upon information provided by the UCSD vice chancellor).

322. *See* LAVERGNE & WALKER, *supra* note 306, at 4 tbl.II; Holley & Spencer, *supra* note 32, at 252 tbl.1.

323. *See* California Resident Freshmen, *supra* note 321.

324. *See* Jesse McKinnon, *The Black Population: 2000* (2001), U.S. CENSUS BUREAU REPORT, *available at* http://www.census.gov/population/www/cen2000/briefs.html (last visited June 25, 2002).

325. *See* Johnson v. Bd. of Regents of the Univ. System of Georgia, 263 F.3d 1234 (11th Cir. 2001).

326. *See* Brief on Appeal of Intervenors Antoine Hester et al. at 17, Johnson v. Board of Regents of the Univ. System of Georgia, 263 F.3d 1234 (11th Cir. 2001) (reporting that African Americans were 243 of 4,272 freshmen in 1999 and 246 of 4,244 freshmen in 1997). The Intervenors in *Johnson* were represented by the NAACP Legal Defense and Educational Fund. *See id.*

than 6% of the point total.[327] When the University of Georgia discontinued its affirmative action plan in 2001 as it proceeded with its appeal, it still adhered tightly to this traditional SAT/GPA definition of merit for the vast majority of admissions decisions, and African American freshmen enrollments dropped by one quarter.[328]

These stark statistical disparities in California, Texas, and Georgia bring us full circle to Professor Lawrence's observation at the beginning of this article that the end of affirmative action is a reminder, for those who need to be reminded, that racial privilege in America based upon Whiteness is alive and well.[329] In this article, we attempted to identify and analyze one important expression of that privilege: racial bias on standardized tests like the SAT. While higher education inequities and standardized test score differences undoubtedly stem from a number of social forces—residential/educational segregation's contribution to inferior K-12 schooling for students of color is a salient example[330]—we argue that the SAT also creates "built-in headwinds" in its own right. We combined empirical evidence with a review of the

---

327. *See* Johnson, 263 F.3d at 1240-41 (reporting that race was 0.5 points out of a maximum of 8.5 points for applicants given further consideration after the bulk of applicants were admitted or rejected automatically based on index scores); Press Release, University of Georgia, Nov. 9, 2001 (reporting that 80-90% of admissions in recent years were based solely on grades and SAT/ACT scores), *available at* http://www.usg.edu/news/2001/11.09.01.html (last visited June 25, 2002).

328. *See* Janet L. Conley, *Race Matters: Michigan Case Reopens Issue in Admissions, Enrollment of Black Freshmen at UGa Declined to Less than 5 Percent in 2001*, FULTON CO. DAILY REPORT, May 23, 2002 (reporting a 24% drop, from 256 African Americans in 2000 to 207 in 2001); Joan Stroer, *UGa's Black Enrollment Holds Steady*, FLORIDA TIMES-UNION, Aug. 18, 2001, at B1 (reporting a one-year drop in African American freshmen enrollments from 249 to 201 based on preliminary data); Sara Hebel, *U. of Georgia Eliminates Use of Race in Admission Decisions*, CHRON. HIGHER EDUC. Dec. 14, 2001, at A26 (reporting that except for athletes and a few dozen students with special skills, admission decisions at the University of Georgia would be based upon high school GPA in core courses and standardized test scores).

329. *See supra* note 1 and accompanying text.

330. On segregation and related educational inequality issues, see, e.g., William D. Henderson, *Demography and Desegregation in the Cleveland Public Schools: Toward a Comprehensive Theory of Educational Failure and Success*, 26 N.Y.U. REV. L. & SOC. CHANGE 457 (2000-2001); Denise C. Morgan, *The New School Finance Litigation: Acknowledging That Race Discrimination in Public Education is More than Just a Tort*, 96 NW. L. REV. 99 (2001); LEONARD S. RUBINOWITZ & JAMES E. ROSENBAUM, CROSSING THE CLASS AND COLOR LINES: FROM PUBLIC HOUSING TO WHITE SUBURBIA (2000); James E. Ryan, *Schools, Race, and Money*, 109 YALE L.J. 249, 257 (1999); Gary Orfield, *Toward an Integrated Future: New Directions for Courts, Educators, Civil Rights Groups, Policymakers, and Scholars*, *in* DISMANTLING DESEGREGATION: THE QUIET REVERSAL OF BROWN V. BOARD OF EDUCATION 331-61 (Gary Orfield et al. eds., 1996); Wendy Parker, *The Supreme Court and Public Law Remedies: A Tale of Two Kansas Cities*, 50 HASTINGS L.J. 475 (1999); Sharon Elizabeth Rush, *The Heart of Equal Protection: Education and Race*, 23 N.Y.U. REV. L. & SOC. CHANGE 1 (1997).

educational literature to argue that the SAT's test construction process unintentionally exacerbates the disparate impact of the test. The problems we have identified will in no way be rectified by ETS's proposed changes to the SAT scheduled for 2005.[331] Moreover, test assembly/item construction is only one manifestation of racial bias in standardized testing that has not garnered sufficient attention, yet is alarming in creating disparate impact.[332]

If the SAT contains racial bias, the question remains where should American higher education go from here? Certainly affirmative action programs can help to counteract the negative impact of racial bias in standardized tests, as the interveners in *Grutter* argued in defending the program at the University of Michigan Law School.[333] In addition, we argued in this article that the SAT test construction process can be altered to decrease the disparate impact of the test on African Americans and Chicanos. While the majority of psychometricians would most likely disfavor our recommended changes, we should point out that our position is not entirely outside of the mainstream. For example, in the *Standards for Educational and Psychological Testing*, jointly produced by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, standard 7.11 states that it can be appropriate to take account of disparate impact: "[w]hen a construct can be measured in different ways that are approximately equal in their

331. *See* Eric Hoover, *College Board Approves Major Changes for the SAT*, CHRON. HIGHER EDUC., June 28, 2002; Tanya Schevitz, *SATs Gain an Essay, Lose the Analogies*, S.F. CHRON., June 28, 2002, at A3.

332. Other forms of test bias such as "stereotype threat" were not covered in this article. For an overview of the stereotype threat literature, see Clark D. Cunningham et al., *Passing Strict Scrutiny: Using Social Science to Design Affirmative Action Programs,* 90 GEO. L.J. 835, 839 (2002) (summarizing stereotype threat research and concluding, "[S]tereotype threat theory is now widely accepted within the field of psychology"); William C. Kidder, *Does the LSAT Mirror or Magnify Racial and Ethnic Differences in Educational Attainment? A Study of Equally Achieving "Elite" College Students*, 89 CAL. L. REV. 1055, 1085-89 (2001) (summarizing several stereotype threat studies). For more detailed research, see, e.g., Jim Blascovich et al., *African Americans and High Blood Pressure: The Role of Stereotype Threat,* 12 PSYCHOL. SCI. 225 (2001); Claude M. Steele, *Thin Ice: "Stereotype Threat" and Black College Students*, ATLANTIC MONTHLY, Aug. 1999, at 44; Steven J. Spencer et al., *Stereotype Threat and Women's Math Performance*, 35 J. EXPERIMENTAL SOC. PSYCHOL. 4 (1999); Joshua Aronson et al., *When White Men Can't Do Math: Necessary and Sufficient Factors in Stereotype Threat*, 35 J. EXPERIMENTAL SOC. PSYCHOL. 29 (1999); Claude M. Steele, *A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance*, 52 AM. PSYCHOL. 613 (1997); Claude M. Steele & Joshua Aronson, *Stereotype Threat and the Intellectual Test Performance of African Americans*, 69 J. PERSONALITY & SOC. PSYCHOL. 797 (1995), *reprinted in* THE BLACK-WHITE TEST SCORE GAP 401 (Christopher Jencks & Meredith Philips eds., 1998).

333. *See supra* Section I.A.

degree of construct representation and freedom from construct-irrelevant variance, evidence of mean scored differences across relevant subgroups of examinees should be considered in deciding which test to use."[334]

We anticipate that opponents of our recommended SAT item bias reduction procedures will criticize us for advocating "race-norming" in the test assembly process.[335] We conclude by reminding readers that, based on our empirical findings and review of the educational measurement literature, the process currently used to construct the SAT, LSAT, GRE, and similar tests unintentionally operates to select questions with larger racial and ethnic disparities (favoring Whites). Thus, the argument that lessening disparate impact in SAT test assembly amounts to unfair racial gerrymandering ignores the current manner in which standardized tests are developed—which incorporates significant behind-the-scenes racial gerrymandering. We believe that the costs of reifying this status quo standardized testing regime (a system that is far from "race-neutral") are too high for America's educational future, particularly for students of color.

Critics of our proposed changes on the SAT will likely argue that modifying the test assembly process to take cognizance of item impact will degrade the predictive validity of the SAT. Yet, for mathematical reasons having to do with the relationship between reliability and predictive validity, ETS researchers such as Stocking acknowledge that "substantial room" exists to lessen disparate impact without compromising the SAT's ability to predict college grades; indeed, by removing construct-irrelevant variance associated with race and ethnicity, the changes we advocate may even create the "win-win" situation of a *less* biased SAT that has *higher* predictive validity than the current form.[336]

---

334. AM. EDUC. RESEARCH ASS'N ET AL., STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING 83 (1999).

335. *Cf.* Roger Clegg, *The Right Score?: The Taint of Race-Norming is Just One Flaw in the Proposed 'Strivers' Rating for SAT-Takers*, LEGAL TIMES, Sept. 20, 1999, at 19-20; Abigail Thernstrom, *The End of Meritocracy: Should the SAT Account for Race? Opposing Opinions by Nathan Glazer and Abigail Thernstrom*, NEW REPUBLIC, Sept. 27, 1999, at 26, *available at* http://www.tnr.com/archive/0999/092799/coverstory092799.html (last visited July 1, 2002); Shelby Steele, *We Shall Overcome—But Only Through Merit*, WALL ST. J., Sept. 16, 1999, at A30; Linda S. Gottfredson, *Racially Gerrymandering the Content of the Police Tests to Satisfy the U.S. Justice Department: A Case Study*, 2 PSYCHOL. PUB. POL'Y, & L. 418 (1996).

336. *See* Stocking et al., *supra* note 124. In this study that attempted to simultaneously reduce race and gender impact, Stocking et al. conclude:

Finally, as a pragmatic matter ETS and the College Board are very unlikely to adopt impact reduction techniques in connection with the SAT unless outside pressure is so substantial as to impact the SAT marketplace. ETS has known about the feasibility of *Golden Rule*-style test assembly procedures for two decades, yet it has only sporadically conducted experimental research on the question—rather than putting something into place on a real SAT.[337] The political difficulties involved are also apparent in the way that ETS quickly retreated from its "Strivers" research—which investigated the development of a scale adjusting SAT scores depending on the sociological obstacles students encountered—immediately after a story appeared in the *Wall Street Journal* and critical op-ed pieces started popping up nationwide.[338] Thus, it may be practical to focus energy on urging colleges and universities either to not use the SAT or at least give applicants the choice of whether or not they want it to be considered in admissions decisions.[339]

> The predictive validity of the SAT I Mathematical, when corrected both for restriction of range and the unreliability of the criterion of first-year grade point averages, is .53. The reliability of a test cannot be less the square of the predictive validity. (This is the inverse of the more familiar statement that predictive validity cannot be greater than the square-root of the reliability.) Thus the reliability of the SAT I Mathematical cannot be less than .28 (.53 * .53) without lowering the predictive validity. Because the current reliability of different editions of the SAT I Mathematical is typically above .90, there is substantial room for a reduction in reliability (.90 minus .28) before predictive validity is constrained by this mathematical relationship. Therefore, it is unlikely that reductions of reliability caused by the approach to test construction used in this paper will constrain predictive validity, and, as demonstrated above, predictive validity is most likely to be increased by this approach.

Id. at 44-45 (citations omitted).

We are confident that impact reduction will not meaningfully decrease the predictive validity of the SAT. However, Stocking et al.'s argument about *improving* the SAT's predictive validity is less certain, for it relies upon assumptions not only about the SAT but about the adequacy and fairness of the criterion variable (college freshman grades). If there is race-related construct irrelevant variance (bias) that is unfortunately common to both the predictor and the criterion, then its removal from the predictor alone would not boost predictive validity.

337. *See supra* Part III.C.

338. *See* Nicholas Lemann, *Tinkering with the Test*, N.Y. TIMES, Sept. 13, 1999, at A19; Ben Gose, *More Points for 'Strivers': the New Affirmative Action?*, CHRON. HIGHER EDUC., Sept. 17, 1999, at A55; Claire Barliant, *Striving to Stay Alive*, SALON.COM, Oct. 18, 1999, *at* http://www.salon.com/books/it/1999/10/18/strivers (last visited Dec. 28, 2001).

339. *See supra* Part IV.D.

***